Konstruktrepräsentation durch regelgeleitete Itemkonstruktion

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Universität Düsseldorf

vorgelegt von
Michael Habon
aus Duisburg

1987

UNIV.
LEIDEN
BIBL.

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41543909_0012

Konstruktrepräsentation durch regelgeleitete Itemkonstruktion

Inaugural - Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Universität Düsseldorf

vorgelegt von

Michael Habon

aus Duisburg

Fotodruck Mainz GmbH, Aachen

1987

D 61

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Universität Düsseldorf

Referent: Prof. Dr. Hornke

Korreferent: Prof. Dr. Manz

Tag der mündlichen Prüfung: 4. Juni 1987

Inhalt

		Seite
1.	Einleitung	1
2.	Validitätskritische Anforderungen an Aufgabenmengen in der pädagogischen und psychologischen Testdiagnostik	9
2.1	Aspekte der Kontentvalidität	9
2.2	Aspekte der Konstruktvalidität	18
2.3	Konzepte interner Testvalidierung	24
2.3.1	STERNBERG's "componential analysis"	24
2.3.2	"Construct representation" sensu EMBRETSON	40
3.	Anforderungen an formale Modelle zur Repräsentation latenter Verhaltensdispositionen ...	45
3.1	Spezifische Objektivität	52
4.	Logistische Aufgabenbearbeitungsmodelle	63
4.1	Die logistischen Komponentenmodelle von SPADA und EMBRETSON (WHITELY)	63
4.2	Das Lineare Logistische Testmodell	73
4.2.1	Der Operationsbegriff im LLTM	76
4.2.2	Entstehungszusammenhänge von Hypothesen über kognitive Lösungsoperationen	78
5.	Überlegungen zu Parameterschätzungen und Gültigkeitskontrollen beim RM und LLTM	81
5.1	Verfahren zur Schätzung der Parameter des RM	81
5.2	Die Schätzung der Operationsparameter	87
5.3	Modelltests für das RM und das LLTM	90
6.	Konkretisierung der Problemstellung	103

Seite

7. Die Modellierung kognitiver Lösungsoperationen bei Matrizenaufgaben 105
7.1 Die Untersuchungen von FORMANN und NÄHRER 105
7.2 Weitere Untersuchungen zur Entwicklung eines Konstruktionsrationals für algorithmische Denkaufgaben 109
7.3. Erweiterung und Revision des Konstruktionsmodells 116
7.3.1 Das Beschreibungssystem von JACOBS und VANDEVENTER 121
7.3.2 Das Klassifikationssystem von WARD und FITZPATRICK 123
7.3.3 Die Beschreibung von PUTZ - OSTERLOH 125
7.3.4 Die Erweiterung des Konstruktionsansatzes von FORMANN 127
7.3.5 Die Revision des Regelsatzes von FORMANN und NÄHRER 133
7.3.6 Eine Datenanalyse auf der Grundlage des revidierten Operationsinventars 137
7.4 Ein Rational zur Konstruktion der Wahlantworten ... 140

8. Die Konstruktion des Itempools 148

9. Die Vernetzung der Testversionen 151

10. Testmaterial und Datenerhebung 158

11. Ergebnisse - Darstellung und Diskussion 161
11.1 Gültigkeitskontrollen zum RM 161
11.1.1 Die Zusammenstellung homogener Probandenstichproben 161
11.1.2 Die Zusammenstellung homogener Aufgabenteilmengen 167

Seite

11.1.3 Linkfit - Analyse und Überblick
über die Kontrollen zum RM 180
11.2 Die Zusammenstellung der Commonscale 185
11.3 Die Problematik inhomogener Items 187
11.4 Gültigkeitskontrollen zum LLTM 202

12. Eine Kreuzvalidierungsstudie 216
12.1 Design und Durchführung 220
12.2 Ergebnisse - Darstellung und Diskussion 223

13. Schlußbemerkungen 230

14. Literaturverzeichnis 234

Anhang

1. Einleitung

Vor nunmehr etwa einem Jahrzehnt erschienen die ersten deutschsprachigen Publikationen, die damalige Perspektiven einer sich gerade abzeichnenden "Computerrevolution" im Hinblick auf das pädagogische und psychologische Testen diskutierten (BUERLI, 1972, 1973; HORNKE, 1976,1977). Der vielerorts eingeschränkte Zugriff auf zentrale Rechnerkapazitäten beschränkte den Einsatz des Computers praktisch auf Testauswertungen, die statistische Behandlung von Ergebnissen und bestenfalls auf eine maschinisierte Aufnahme von Testdaten mittels Markierungslesern. Dennoch hoffte man, die schnell wachsende Verfügbarkeit über Rechner- und Speicherkapazitäten bald nutzen zu können, um einen schon länger, vor allem in den USA, diskutierten INTERAKTIVEN TESTPROZESS (LORD, 1970; GREEN, 1970; HOLTZMAN, 1971; WOOD, 1973) zu realisieren, der einige fundamentale, im folgenden kurz skizzierte, Probleme herkömmlicher Testdiagnostik zu lösen versprach.

Traditionelle Testverfahren bestanden -und bestehen- abgesehen von wenigen Ausnahmen, wie z.B. den Binetarien oder dem Hetzer-Test, aus einer begrenzten, feststehenden Zahl von Aufgaben, die, wenn sie einmal die Hürden einer i.d.R. einmaligen empirischen Analyse mehr oder minder erfolgreich genommen hatten, in einer starren Reihenfolge als Einzel- oder Gruppentests administriert werden.

In der traditionellen Testkonstruktionslehre sensu GULLIKSEN (1950) oder LIENERT (1969,3) ist das Augenmerk auf die Konstruktion von Tests gerichtet, wobei das Reliabilitätskonzept und die abgeleiteten Itemselektionstechniken der als "klassisch" bezeichneten Testtheorie die Etablierung von Tests begünstigen, deren Aufgaben in einer definierten Population hohe Beantwortungsvarianz aufweisen. Man kann wohl davon ausgehen, daß die Überrepräsentation von Aufgaben "mittlerer" relativer Lösungs-

häufigkeit in solchen Tests oftmals hinreichend präzise Aussagen über die Merkmalsausprägung von Personen zuläßt, für die solche Items "mittelschwer" sind, d.h. mit anderen Worten, deren Rohwert (Score) im mittleren Bereich einer durch die Testwerte definierten Rangordnung liegt. Bei unterstellter Normalverteilung der Testwerte sind reliable "Messungen" für den größeren Teil der durch die Eichstichprobe definierten Population möglich. Der traditionell konstruierte Test ist in der Regel ein an den Durchschnitt der Leistungsfähigkeit einer Population adaptiertes Differenzierungsinstrument (vgl. LORD, 1970). Viele solcher Tests wurden und werden aber in Situationen eingesetzt, in denen sie aufgrund ihrer Schwierigkeitszusammensetzung keine reliablen Informationen liefern können, wie z.B. bei der Identifikation von Hochbegabten oder der Diagnose defizitärer Lernkapazitäten. Dies ist allerdings weniger vielleicht allzu unkritischen Testanwendern, als vielmehr einem ungeeigneten Reliabilitätskonzept anzulasten, das die Fidelität der Messung auch dann vortäuscht, wenn für bestimmte Personen die Mehrzahl der Testaufgaben entweder zu "leicht" oder zu "schwer" war (vgl. FISCHER, 1974, S.114ff.).

Analysen konventioneller Tests mit probabilistischen (Item Response-) Theorien (IRT; LORD & NOVICK, 1968; FISCHER, 1974), die von einem je Testwert verschiedenen Meßfehler ausgehen, ließen die Unzuverlässigkeit der Messungen in diagnostisch oftmals interessierenden Randbereichen der Skalen sg. "peaked" Tests (LORD, 1980) als dringend zu lösendes Problem bewußt werden. Grundsätzlich ist diesem Mangel auch dann mit konventionellen Tests nicht wirkungsvoll zu begegnen, wenn sie sich aus wenigen Aufgaben zusammensetzen, die gleichmäßig über ein breites Schwierigkeitsspektrum verteilt sind. In einem solchen Fall stehen an diagnostisch möglicherweise relevanten Skalenpunkten nur sehr wenige informationsreiche Aufgaben bereit. Dies führt zu einer relativ gleichmäßigen, jedoch niedrigen Meßgenauigkeit in allen Skalenbereichen.

Der Konstrukteur eines konventionellen Tests sieht sich somit stets mit einem unauflösbaren "Bandbreiten - Fidelitätsdilemma" (CRONBACH & GLESER, 1965, S.97; Übersetzung des Verfassers) konfrontiert. Diese Einsicht motivierte die Suche nach alternativen "Teststrategien" (vgl. HORNKE, 1976), die allgemein darauf abzielen, Aufgaben je nach diagnostischer Fragestellung zusammenzustellen und schließlich darin gipfeln dem Probanden (PB) in einer interaktiven -computerunterstützten- Testprozedur nur solche Items vorzulegen, die seinem Fähigkeits-, Leistungs- oder Einstellungsniveau angemessen, also nicht "zu schwer" oder "zu leicht" sind. Ziel eines derart konsequent durchgeführten adaptiv antwortabhängigen ("individualized-", FISCHER & PENDL, 1980; "tailored-", LORD, 1980; "response contingent-", WOOD, 1973) Testens sind auftragsbezogen präzise Messungen im interessierenden Range des fraglichen latenten Merkmals bei gleichzeitig maximaler Effizienz/ Ökonomie des Verfahrens und minimaler Belastung bzw. optimaler Motivation des Probanden (vgl. WEISS, 1982).

Natürlich stellen dazu erforderliche Personen- und Konfidenzintervallschätzungen, die Verwaltung der Items und ihre Präsentation in einer adaptiven Testprozedur nicht gerade geringe Anforderungen an Hard- und Software. Der technologische Entwicklungsstand setzte noch vor rund zehn Jahren der Realisation und Verbreitung interaktiver computerunterstützter Teststrategien außerhalb experimenteller Umgebungen enge Grenzen. Während in der amerikanischen Literatur zum adaptiven Testen (LORD, 1970; GREEN, 1970; HOLTZMAN, 1971; WOOD, 1973) aber schon durchgängig von der Annahme ausgegangen wurde, daß solche Verfahren einmal computerunterstützt durchgeführt werden, wurde in der deutschsprachigen Literatur dieser Problemkreis nur selten diskutiert (HORNKE, 1975) oder wegen der vermuteten "relativ geringen praktischen Bedeutung" (FISCHER, 1974, S.337f) gänzlich ausgekammert.

Heute sind mit schnellen, flexiblen und portablen Kleincomputern der PC-Klasse Technologien verfügbar, die computerunterstützte

UNIV. LEIDEN PSYCH.

adaptive Teststrategien (CAT) auf der Grundlage gängiger Latent Trait Modelle AUS TECHNISCHER SICHT ermöglichen (vgl. BAKER, 1984).

In der Zwischenzeit hatte man jedoch erkannt, daß CAT das "Bandbreiten - Fidelitätsdilemma" nur dann mindern oder gar lösen kann, wenn in einer homogenen Itemmenge für ALLE relevanten Skalenabschnitte ausreichend viele geeignete Aufgaben verfügbar sind (vgl. z.B. LORD, 1977; JENSEMA, 1977). Nur dann scheint es möglich, Probanden (Pbn) aller Merkmalsniveaus Items angemessener Schwierigkeit zu präsentieren, um sie so mit jeweils zunehmender statistischer Sicherheit auf der zugrundeliegenden Skala zu lokalisieren.

Gegen bloße "Computerisierungen" konventioneller Tests spricht aber nicht allein, daß potentielle Möglichkeiten adaptiver Strategien zwangsläufig auf denkbare "Verkürzungen" konventioneller Verfahren für bestimmte Probandengruppen reduziert sind, sondern insbesondere auch, daß die Interpretierbarkeit der unterstellten Skalen wegen gänzlich fehlender oder unzureichender inhaltlicher oder meßtheoretischer Fundierung zumeist fraglich ist (vgl. z.B. MCNEMAR, 1964; ANASTASI, 1967; BOCK & WOOD, 1971; MESSICK, 1972; FISCHER, 1974; GLASER & PELLEGRINO, 1978).

Nahezu alle derzeit gebräuchlichen Tests basieren auf traditionellen Konstruktionsprinzipien (vgl. LORD & NOVICK, 1968; LIENERT, 1969/3; THORNDIKE, 1982). Die ihnen zugrundeliegende sog. "klassische" Testtheorie macht nur marginal Aussagen über das Item; Gegenstand der Evaluation ist vorrangig die formale Revision einer zumeist nach intuitiven Vorverständnissen erstellten Aufgabensammlung als Ganzes. Dabei werden zentrale Validitätsfragen im Rahmen nachgelagerter Konstruktvalidierungsstudien meist mit Bezug auf den Gesamttest diskutiert. Beim computerunterstützten adaptiven Testen rückt dagegen das Item schon eher als "Meßinstrument" in den Mittelpunkt der Betrach-

tungen. Um sinnvolle, d.h. inhaltlich interpretierbare Aussagen über individuelle Merkmalsausprägungen zu ermöglichen, ist für jede Aufgabe sicherzustellen, daß sie das fragliche Merkmal repräsentiert, also valide ist, und daß sie das gleiche Merkmal erfaßt wie alle anderen Aufgaben, also Homogenität der Skala gewährleistet ist. Beide Kriterien können durch die klassische Testtheorie aber nur unzureichend gewährleistet werden. Der Hauptgrund dürfte wohl darin bestehen, daß die Konstruktion der Items nicht Bestandteil dieser formalstatistischen und damit inhaltsleeren Theorie von Testwerten ist (vgl. SPADA & FREY, 1973; HILKE, 1980).

Charakteristisch ist, daß einschlägige Lehrbücher zur Testkonstruktion (z.B. LIENERT, 1969/3) Modellen und Methoden der Testevaluation breiten Raum geben, während sie der vorgelagerten inhaltlichen Repräsentation des operationalisierten Merkmals in den Testitems durch Hinweise auf Itemformate und -formulierungen gerecht zu werden versuchen, gerade so, als ob sich allein damit schon Items konstruieren ließen, die ein Merkmal in der beabsichtigten Weise differenziert erfassen. Es verwundert daher kaum, daß die gängige Praxis der Itemkonstruktion gelegentlich als "Kunst" aufgefaßt wird, die sich dem theoretisch fundierten Zugriff weitgehend entzieht. So schreibt WESMAN (1971, S.81):

> "Item writing is essentially creative - it is an art. Just as there can be no set of formulas for producing a good story or a good painting, so there can be no set of rules that guarentees the production of good test items. Priciples can be established and suggestions offered, but it is the writers judgement in the application -and occasional disregard- of theses principles and suggestions that determines whether good items or mediocre ones are produced. Each item, as it is being written, presents new problems and new opportuni-

> ties. Thus item writing requires an uncommon combination of special abilities and is mastered only through extensive and critically supervised practice."

Die Frage, welche Merkmale wie meßbar sind, bleibt im Rahmen traditioneller psychologischer Testkonstruktion ungestellt und damit unbeantwortet. Vielmehr wird versucht, einer informell zusammengestellten Itemmenge a posteriori durch konstrukt- oder kriterienbezogene Validierungsstudien eine bestimmte Gültigkeit anzutragen, indem die Beziehungen des zu anderen, auf gleiche Weise entstandenen Instrumenten bzw. zu Kriterienmaßen, deren Validität ihrerseits meist nur per fiat angenommen werden kann, zumeist mit i.w.S. korrelationsstatistischen Techniken untersucht wird. Gerade in der Intelligenzforschung versuchte man jahrzehntelang durch explorative Faktorenanalysen festzustellen, welche inhaltlichen Aspekte intellektueller Fähigkeiten vorhandene Itemsammlungen operational definieren. An anderer Stelle wurde schon gezeigt (vgl. z.B. das "Münchener Symposium über Faktorenanalyse", 1966; FISCHER, 1974), daß die Faktorenanalyse als deskriptive statistische Methode der Datenreduktion diesen Anspruch schon aus formalen Gründen nicht einlösen kann. Daß sie ihn de facto nicht einzulösen vermochte, wird schon daran deutlich, daß am (vorläufigen) Ende einer relativ unsystematischen Suche nach den "Dimensionen der Intelligenz" eine Vielzahl hierarchischer (SPEARMAN, BURT, VERNON, CATTELL) und struktureller (THURSTONE, MEILI, GUILFORD) "Faktorenmodelle" der Intelligenz stehen, von denen jedoch kaum eines befriedigend empirisch abgesichert ist (vgl. z.B. HORN, 1968; HORN & KNAPP, 1973), geschweige denn hinreichend expliziert, auf welche Klasse von Problemlösungsvorgängen die "gemessenen" Fähigkeiten verallgemeinerbar sind (vgl. PUTZ-OSTERLOH, 1981; DÖRNER, 1984).

Erwartet man von einem Meßinstrument, daß es meßtheoretischen und validitätskritischen Anforderungen genügt, dann besteht bereits bei der Aufgabenkonstruktion die Notwendigkeit pädagogisches bzw. psychologisches Merkmalswissen konsequent in Aufgabenkonstruktion(sstrategien) umzusetzen.

In der vorliegenden Arbeit soll versucht werden bisher nur global angesprochene Anforderungen an Aufgabenmengen u.a. in Abhängigkeit von ihren Verwendungszwecken in der pädagogischen und psychologischen Testdiagnostik zu spezifizieren. Wenn hier zwischen pädagogischer und psychologischer Testdiagnostik unterschieden wird, so allein, um graduell verschiedenen Testzielen und -zwecken Rechnung zu tragen. Während es in der pädagogischen Testdiagnostik vorrangig um die Erfassung von Variablen geht, "die durch Definition auf einen präzisierten VERHALTENSbereich bezogen (sind)" (KLAUER, 1984, S.693ff., Hervorhebung vom Verfasser), steht in der psychologischen Testdiagnostik traditionell die Messung LATENTER Variablen im Vordergrund. Infolgedessen orientieren sich Testkonstruktionen schwerpunktmäßig an verschiedenen Validitätsbegriffen, nämlich der "Kontentvalidität" (vgl. KLAUER, 1978) bzw. an der "Konstruktvalidität" (vgl. MESSICK, 1975) der Instrumente.

Um die Anforderungen einer meßzielorientierten Aufgabenkonstruktion zu präzisieren, sind Explikationen beider Validitätsbegriffe auf konzeptuelle Überschneidungen und Spezifitäten zu hinterfragen. Aus einer Redefinition des Begriffs der Konstruktvalidität, die Aspekte der Kontentvalidität einschließt, wird die Notwendigkeit einer empirisch prüfbaren, formalen Aufgabenbeantwortungstheorie als Voraussetzung für die rationale Konstruktion einer Itembank zur (adaptiven) Messung einer latenten Fähigkeit abgeleitet.

Im zweiten Teil der Arbeit wird die Entwicklung und Validierung einer formalen Lösungstheorie für einen bestimmten strukturellen

Aufgabentyp (Matrizenitems) zusammengefaßt und methodische Voraussetzungen des Designs und der Evaluation einer Itembank für das computerunterstützte Testen einer latenten algorithmischen Problemlösungsfähigkeit diskutiert. Anschließend werden die Ergebnisse eines Itembankkonstruktionsexperiments dargestellt und unter der Fragestellung diskutiert, ob und inwiefern sie Evidenzen für eine an Rationalen orientierte Aufgabenkonstruktion erbringen.

2. Validitätskritische Anforderungen an Aufgabenmengen in der pädagogischen und psychologischen Testdiagnostik

2.1 Aspekte der Kontentvalidität

Nach einer systematischen Übersicht über verschiedene typische Entscheidungssituationen in der pädagogischen und psychologischen Diagnostik kommen CRONBACH & GLESER (1965/2) zu dem Schluß, daß die Klassische Testtheorie (KTT) von 64 potentiellen Entscheidungssituationen lediglich eine behandelt. Das Fehlen einer übergeordneten Testtheorie führe dazu, daß das Entscheidungsproblem häufig erst gar nicht spezifiziert und die Testkonstruktionstechniken der für den spezifischen Fall unangemessenen KTT oft unreflektiert übernommen werden. Dies führte insbesondere im Bereich pädagogischer Fragestellungen zu Konsequenzen, die ursprünglich nicht intendiert waren. So stellt WARD (1970, S.315) treffend fest, daß

> "the orthodox psychometrist (..is..) looking anxiously for facility values of 50 per cent, whilst the educationalist who has expended much time and effort on a basic instructional programm may be hoping for values of 100 per cent."

Im allgemeinen Zusammenhang mit pädagogisch-psychologischen Fragestellungen der Erfolgskontrolle von Unterricht (vgl. SCHOTT, 1983), insbesondere aber im Zuge einer sich langsam entwickelnden Theorie der programmierten Unterweisung und der resultierenden Notwendigkeit ihre Effizienz zu bewerten, begann mit den Artikeln von EBEL (1962) und GLASER (1963) die Diskussion um sg. "criterion-referenced-" (GLASER & NITKO, 1971) oder "lehrzielorientierte" Tests (FRICKE, 1974). Sie sollten geeignet sein festzustellen, ob und ggfs. wie gut ein inhaltlich definiertes Lehrziel/ Kriterium erreicht ist. Die verwendeten Testaufgaben sollen dabei das Lehrziel repräsentieren und in erster Linie dazu dienen, die individuelle Leistung eines Schülers mit einer

"idealen" Norm zu vergleichen (vgl. FRICKE, 1974). Die bei solchen Vergleichen auftretenden test- und entscheidungstheoretischen Probleme interessieren hier weniger als die Voraussetzungen, die Aufgabenmengen erfüllen sollten, um die von EBEL (1962) geforderte DIREKTE INTERPRETIERBARKEIT des Testverhaltens zu ermöglichen.

In der kriteriumsorientierten Leistungsmessung wird, ähnlich wie in der jüngeren experimentellen Denkpsychologie (NEWELL & SIMON, 1972; DÖRNER, 1976) Verhalten als Änderung von Zuständen dargestellt.

> "Der Anfangszustand einer Aufgabe beschreibt ihren Inhaltsaspekt vor der Aufgabenlösung (...), der Endzustand den Inhaltsaspekt nach der Aufgabenlösung (...), die Änderung bei der Aufgabenlösung besteht in der Überführung vom Anfangs- zum Endzustand und betrifft den Verhaltensaspekt (...). Verhalten wird hier also analog dargestellt wie die extensionale Beschreibung einer Funktion durch die Angabe der beiden Mengen, die aufeinander abgebildet werden." (SCHOTT & WIEBERG, 1984, S.54)

Die Änderung eines bestimmten Anfangszustandes in einen bestimmten Endzustand beschreibt eine "Aufgabenklasse". Ihre vollständige Definition wird somit als hinreichende Beschreibung des interessierenden Testverhaltens aufgefaßt. Objektivierbare Stimulus- und Reaktionskomponenten stellen dabei, vom Behaviorismus beeinflußt, die definitorischen Kategorien für Verhalten(-sbereiche) bzw. Aufgabenmengen dar.

Die Stimuluskonfiguration, die dem Probanden (Pb) in Form der Testaufgabe begegnet, ist nach KLAUER (1984) vollständig beschrieben, wenn sowohl der eigentliche inhaltliche Aspekt der Aufgabe (=Sachverhalt), als auch die Art und Weise, WIE der Sachverhalt vorgelegt wird, eindeutig festgelegt sind. Mit dem

"wie" ist sowohl die Aufgabenform (Aufgabentyp) als auch der übrige, in der Regel standardisierte situative Kontext der Testdurchführung (Instruktion, Administration als Einzel- oder Gruppentest etc) angesprochen. Wie verschiedene Sachverhalte durch Aussagen, Aussageformen oder Flußdiagramme objektiv beschrieben werden können, wurde bereits mehrfach zusammenfassend dargestellt, worauf an dieser Stelle verwiesen sei (vgl. KLAUER, 1978, 1984; ROID & HALADYNA, 1982; HORNKE & HABON, 1984b).

Die Responsekomponente einer Aufgabe bzw. Aufgabenmenge ist durch den gewählten Aufgabentyp ("item form"; OSBURN, 1968) beschrieben. Er bzw. sie allein entscheidet, welche "offenen", beobachtbaren Reaktionen (Auswahlantwort ankreuzen; eine Antwort frei formulieren o.ä.) vom Pb erwartet werden. KLAUER (1984, S.7) argumentiert, daß über den Aufgabentyp in Verbindung mit dem Inhalt die "Anforderungen" an den Pb festgelegt seien, wobei jedoch unklar wäre, welche psychischen Prozesse der Aufgabenbeantwortungen durch die Anforderungen ausgelöst würden. Er weist aber darauf hin, daß in der Literatur Anweisungen für die Gestaltung bzw. Auswahl von Itemformen vorzufinden seien, die unterschiedliche mentale Verarbeitungsprozesse evozieren. In diesem Zusammenhang wird besonders auf die systematischen Ansätze von BLOOM (1956, 1972) sowie KRATHWOHL, BLOOM & MASIA (1975) verwiesen, deren Taxonomien kognitiver bzw. affektiver Lehrziele sowohl Definitionsversuche unterschiedlicher kognitiver Prozeßkategorien, wie "Wissen", "Verstehen" und "Anwenden", als auch Hinweise auf Itemformen enthalten, die geeignet sein sollen, diese Merkmale zu erfassen.

Während in gängigen Verfahren zur Präzisierung von Lehrzielen bzw. Aufgabenmengen genaue Beschreibungen der Verhaltenskomponente durch ausschließlich beobachtbare Verhaltensweisen betont werden (vgl. GLASER und NITKO, 1971), sind mit der Beschreibung des Antwortverhaltens über kognitive oder affektive Komponenten aber nicht beobachtbare hypothetische Konstrukte angesprochen,

die Aussagen darüber implizieren, WIE durch "Wissen" (?), "Verstehen" (?), "Problemlösen" (?) Probanden in definierten Situationsklassen zu bestimmten definierten Antworten (=Reaktionen) gelangen. Die Entscheidung, ob und in welcher Weise "Wissen-", "Verstehens-" oder "Problemlösungsprozesse" an der Beantwortung solchermaßen definierter Aufgaben (-mengen) beteiligt sind, setzt eine empirisch geprüfte bzw. prüfbare Theorie voraus, die diese Kategorien begrifflich isoliert und über objektive Situationskomponenten im o.a. Sinn operationalisiert.

Obgleich die BLOOMschen Taxonomien populär sind und besonders im anglo-amerikanischen Raum angewendet werden, um Lehrpläne zu kategorisieren oder Lehr- und Therapieprogramme zu entwickeln (BLOOM, HASTINGS und MADAUS, 1971; SCHOTT, 1973), fehlen bis heute empirische Evidenzen für die Validität der unterstellten Konstrukte. Operationale Definitionsversuche der Kenntnis, des Verständnisses oder der Anwendung eines verstandenen Sachverhalts durch Aussagen, Aussageformen und Aufgabentypen, gestatten zwar Items zu formulieren, die zweifelsfrei die geforderten Eigenschaften besitzen, nämlich Elemente einer wohldefinierten Itemgrundmenge zu sein, der implizierte Schluß auf UNTERSCHIEDLICHE kognitive Leistungen ist aber so lange unzulässig, wie keine nur einigermaßen zureichenden empirischen Belege für die PSYCHOLOGISCHE Äquivalenz (d.h. hier "Konstruktvalidität") einer solchen Aufgabentypologie gibt.

Vordergründig scheint das Testverhalten direkt inhaltlich interpretierbar und auf das "Universum" interessierender Verhaltensweisen verallgemeinerbar, wenn eine Grundgesamtheit von Items in oben skizzierter Weise über einen Konstruktionsregelsatz definiert und nach bestimmten Auswahlvorschriften (vgl. KLAUER, 1984, S.10f.) eine repräsentative Stichprobe von Testaufgaben zusammengestellt wurde. Itemmengen, die diese Eigenschaften aufweisen, werden als "inhaltlich gültig" oder "KONTENTVALIDE" bezeichnet (vgl. Standards for Educational and Psychological Tests (APA,

1974; KLAUER, 1978)). Dieses "generelle Verbindlichkeit" beanspruchende Konzept von "Kontentvalidität" (KLAUER, 1984, S.1) qualifiziert somit eine objektiv überprüfbare Eigenschaft von Tests, nicht aber von TestWERTEN. In der Literatur zur kriterienorientierten Leistungsmessung und der gerade für den Bereich pädagogischer Diagnostik als essentiell betrachteten Kontentvalidität (vgl. HAMBLETON und NOVICK, 1973), wird aber nahezu durchgängig ein besonderer Vorteil kontentvalider Tests darin gesehen, daß man ohne Rekurs auf latente Dimensionen, Faktoren oder Konstrukte sagen könne, was sie messen. (z.B.: OSBURN, 1968; HIVELY u.a., 1968; GLASER und NITKO, 1971; FRICKE, 1974; HERBIG, 1976; GUION, 1977; KLAUER, 1978).

Gegen diese Suffizienzannahme des Kontentvaliditätskonzepts im Sinne eines Repräsentationsverhältnisses sind verschiedentlich schwerwiegende Einwände erhoben worden (vgl. MESSICK, 1975, 1981; FITZPATRICK, 1983).

> "The major problem here is that content validity (...) is focused upon test FORMS rather than test SCORES, upon INSTRUMENTS rather than MEASUREMENTS. Inferences in education and psychological measurement are made from scores, and scores are a funktion of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself does not provide validity. Call it "content relevance", if you will, or "content representativeness", but do not call it "content validity", because it does not provide evidence for the interpretation of responses or scores" (MESSICK, 1975, S.960f.).

Unter forschungsstrategischen Perspektiven stellt die intersubjektiv nachvollziehbare Definition des Testinhalts eine Möglichkeit dar, Hypothesen über die der Aufgabenbeantwortung zugrunde-

liegenden Konstrukte abzuleiten (vgl. NUSSBAUM, 1980; KLAUER, 1984). Sie ist jedoch keine hinreichende Bedingung für die Interpretierbarkeit der Testwerte als Messungen einer Fähigkeit, Fertigkeit oder Einstellung. Hiermit sind vielmehr Fragen der Kontentvalidität angesprochen, die empirisch zu beantworten sind.

FITZPATRICK (1983) möchte den Validitätsbegriff daher ausschließlich für empirische Testvaliditäten reserviert wissen. Damit folgt sie der Meinung EBELS (1975):

> "Only when one variable is measured in order to make inferences about some presumably related variable, or about some underlying, and hence unobservable determinants of a particular kind of behavior, do real questiones of validity arise (...) content validity is not really a kind of validity at all" (zit. nach GUION, 1977, S.5).

Mit der Abkehr von der in den 60er Jahren vorherrschenden Denkweise des Behaviorismus und Operationalismus, erhebt CRONBACH (1971) die Forderung, das ALLE Tests, also auch kriterienorientierte (pädagogische) Tests neben der Kontentvalidität auch Konstruktvalidität aufweisen müßten.

Auch KLAUER schreibt schon Anfang der 70er Jahre:

> "Es wird endlich an der Zeit, mit der Vorstellung aufzuräumen, bei Schulleistungstests, lehrzielorientierten Tests (oder bei Klassenarbeiten) könne man sinnvoll nur nach der Kontentvalidität fragen" (1972, S.36).

Wenig später erneuert MESSICK (1975, S.957) die Forderung CRONBACHS:

> "... ALL MEASUREMENT SHOULD BE CONSTRUCT - REFERENCED. A measurement estimates how much of something an individual displays or possesses. The basic question is, What is the nature of that something? It may be answered by refering to evidence and support of particular attributes, processes, or traits construed to underlie and determine task performance. A second question is, Is the measurement amount a lot or a little? This may be answered by comparing it to the scores of performance levels of other individuals or groups (which is a norm - referenced answer), or by comparing it to specified levels of minimal, acceptable, or desirable performance (which is a criterion-referenced answer)."

In seiner jüngsten Arbeit zur Kontentvalidität greift KLAUER (1984) die Kontroverse um die Relevanz des Konzepts erneut auf, indem er auf den Verwendungszusammenhang von Tests in der pädagogischen Diagnostik verweist. Psychische Aufgabenverarbeitungsprozesse sind, nach KLAUER, besonders dann von Interesse, wenn es etwa im Rahmen pädagogischer Interventionsprogramme darum geht, spezifische Stärken und Schwächen zu diagnostizieren und entsprechende Förderungsmaßnahmen gezielt einzusetzen. Bei der Zusammenstellung von Aufgabenteilmengen, die bestimmte Verarbeitungsprozesse erfordern, kommt nach Ansicht des Verfassers, generierenden Aufgabenmerkmalen besondere Bedeutung zu, da man auf ihrer Grundlage die Itemmenge in die interessierenden Teilmengen zerlegen könne. Ob diese Teilmengen homogen oder heterogen bezüglich der beanspruchten kognitiven oder affektiven Prozesse sind, ist dann eine Frage der empirischen Konstruktvalidierung. Auf die konzeptuellen und methodischen Probleme, die bei der, in der Regel auf korrelationsstatistischen Verfahren beruhenden Kon-

struktvalidierung im Zusammenhang mit kriterienorientierten Tests entstehen, verweisen CRONBACH (1971) und HILKE (1980).

Im Leistungsbereich, so KLAUER, interessieren weniger die psychischen Prozesse, die bei der Bearbeitung eines kontentvaliden Tests ablaufen, sondern vielmehr das Niveau und Profil der Leistungsfähigkeit. Außerdem wären, wenn es um Stellungnahmen, Wertungen oder ähnliches ginge, Schlüsse auf die Ausprägung der gegenwärtigen Leistung von Bedeutung und weniger die Art und Weise ihres Zustandekommens.

> "Im Leistungsbereich läßt ein kontentvalider Test Schlüsse auf die Fähigkeit des Individuums zu, die Aufgaben der definierten Grundgesamtheit lösen zu können, welche Lösungsprozesse das Individuum dabei auch immer verwendet" (KLAUER, 1984, S.21).

Selbst dieser Versuch, die Kontentvalidität als hinreichendes Validitätskriterium zu legitimieren, scheint unzulässig. So bemerkt MESSICK (1975), daß bei gewährleisteter Kontentrepräsentativität (vgl. FITZPATRICK, 1983), richtige Aufgabenbeantwortungen den Schluß auf eine wie auch immer beschaffene Fähigkeit erlauben mögen, aus Falschbeantwortungen zu schließen, jemand verfüge nicht über die betreffende Fähigkeit, sei jedoch problematisch:

> "The inference of inability or incompetence from the absence of correct performance requires the elimination of a number of plausible rival hypotheses dealing with motivation, attention (...). Thus a report of failure to perform would be valid, but one of inability to perform would not neccecarily be valid. The very use of the term INABILITY invokes constructs of attribute and process, whereas a content-valid interpretation would stick to the outcomes" (S.960).

Zusammenfassend kann festgehalten werden, daß mittels der im Rahmen kriterienorientierter Leistungsmessung entwickelten Techniken zur Itemkonstruktion, wohl definierte Grundmengen von Testaufgaben objektiv erstellbar sind. Definierte Erzeugungsmerkmale können herangezogen werden, um Hypothesen über die der Aufgabenbeantwortung zugrundeliegende Fähigkeit zu generieren. Diese Hypothesen sind dann in einer der Aufgaben- und Testkonstruktionen NACHGELAGERTEN Phase empirisch zu validieren. Die durch das Konzept der Kontentvalidität implizierte direkte Interpretation von Testwerten unter Umgehung einer Konstrukt- oder Kriterienvalidierung ist kurzschlüssig, da sie den differentiellen Einfluß der zwischen objektiven Stimulus- und Responsekomponenten einer Aufgabe vermittelnden spezifischen Verarbeitungsprozesse unberücksichtigt läßt. Genau diese Verarbeitungsprozesse konstituieren aber die Fähigkeit, auf die man auch in der kriterienorientierten Leistungsmessung faktisch generalisiert. Die bloße begriffliche Definition einer Aufgabengrundmenge reicht nicht aus, um resultierende Testwerte im Sinne von Merkmals- oder Fähigkeitsausprägungen zu interpretieren. Sie sind in der Tat bedeutungslos, solange keine empirischen Evidenzen für die Plausibilität einer solchen Testdateninterpretation vorliegen (CRONBACH, 1971). Diese Ansicht teilt auch EMBRETSON (1985) in ihrem jüngst erschienen Buch zum "Test Design":

> "Tests cannot be designed by their substantive properties unless the variables that determine item responses have been explicated. In essence, understanding item responses requires a theory of the variables that are involved in item performance" (S.4).

Die atheoretischen und vor-empirischen "Item-Writing Technologies" der kriterienorientierten Testkonstruktion sensu ROID und HALADYNA (1982) sind damit gleichsam unzureichend, Anforderungen an Aufgabenmengen zu spezifizieren, die potentiell geeignet sind, auf kognitive/ affektive Prozesse, Operationen oder Komponenten (zur Begrifflichkeit vgl. Kp.4.2.1) bezogene Fähigkeiten/ Eigen-

schaften zu erfassen.

2.2 ASPEKTE DER KONSTRUKTVALIDITÄT

Dem Operationalismus der kriterienorientierten Leistungsmessung steht das etwa vor einem Vierteljahrhundert von CRONBACH und MEEHL (1955) formulierte Konzept der KONSTRUKTVALIDITÄT gegenüber.

> "(...) Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is NOT operationally defined" (S. 282).

Zwar gehen schon CORNBACH und MEEHL davon aus, daß die Konstruktvalidierung eines Tests prinzipiell einem Theoriebildungsprozeß entspricht, das zu dieser Zeit noch vorherrschende funktionalistische Theorieverständnis führte aber bekanntlich dazu, daß man weniger an Systemen und Subsystemen intervenierender Aufgabenbeantwortungsprozesse als vielmehr an den konvergenten und divergenten Beziehungen der Variablen in einem theoretisch postulierten "nomologischen" Netz von Variablen interessiert war (vgl. CAMPBELL und FISKE, 1959).

Versuche, die Validität von Tests über ihre (korrelativen) "AUSSENbeziehungen" zu bestimmen, können aber allein deshalb nicht zu einem theoretisch befriedigendem Verständnis der gemessenen Variable führen, da die involvierten Schlußprozesse zirkulär und die resultierenden Validitätsaussagen tautologisch sind, wenn die beteiligten Variablen nicht im o.a. Sinne a priori substanztheoretisch begründet sind.

Dies gilt unbestritten insbesondere dann, wenn das manifeste Testverhalten (= Testwerte) als operationale Variablendefinition angenommen wird ("Intelligence is what the tests test", BORING, 1923,S.187). Im Gegensatz zur Auffassung einiger Testtheoretiker (vgl. z.B. FISCHER, 1974; SCHEIBLECHNER, 1980; MOOSBRUGGER, 1983) trifft dieser Vorbehalt allerdings auch dann zu, wenn das nomologische Netzwerk aus LATENTEN Variablen besteht, die den von LORD und NOVICK (1968) formulierten "minimal hinreichenden Kriterien" für den logischen Status von "Persönlichkeitsmerkmalen" genügen. Danach sind "Persönlichkeitsmerkmale" definiert als "Verhaltensdispositionen", die sich als notwendig (zweckmäßig) für die Verhaltensbeschreibung erweisen und in einer eindeutigen, empirisch grundsätzlich falsifizierbaren Relation zum beobachteten Verhalten stehen" (SCHEIBLECHNER, 1980, S.501). Die potentiell empirisch falsifizierbare Relation zwischen einer hypothetischen latenten Variable oder Verhaltensdisposition und dem manifesten Verhalten wird durch formale, meist statistische Modelle -in mehr oder minder stringenter Form- hergestellt. Auf diesen Problemkreis wird im Zusammenhang mit meßtheoretischen Aspekten der Konstruktrepräsentation näher einzugehen sein (vgl. Kp. 3).

Der empirische Nachweis latenter Dimensionen oder Strukturen gilt als notwendige Voraussetzung, um Testwerte eindeutig als Ausprägungen von Fähigkeiten, Eigenschaften oder Einstellungen interpretieren zu können. Die zentrale Frage, was ein Test mißt, oder genauer, auf welche Klasse von kognitiven bzw. affektiven Verhaltensweisen oder Einstellungen Testwertinterpretationen Bezug nehmen, ist allein durch den Rekurs auf die empirische Geltung gängiger psychometrischer Testmodelle jedoch nicht beantwortbar. Die Behauptung, ein Test messe eine latente Fähigkeit, ist inhaltsleer und keine zweckmäßige Aussage über die Konstruktvalidität eines Tests, wenn sie sich ausschließlich auf die Gültigkeit der durch das jeweilige Testmodell repräsentierten gesetzmäßigen Beziehung zwischen manifestem Testverhalten und formal definierter latenter Verhaltensdisposition bezieht (vgl. HILKE, 1980).

Aus dem vorherrschenden Primat der empirischen Testmodellbewährung gegenüber der inhaltlichen Bestimmung des hypothetischen Konstrukts bzw. der latenten Dimension folgt zudem, daß keine bedeutungsvollen Erklärungen dafür gegeben werden können, WARUM für bestimmte Aufgabengruppen zentrale Annahmen der Testmodelle (z.B. Homogenität/ Eindimensionalität; Itemcharakteristische Kurven; vgl. LORD und NOVICK, 1968) gelten und für andere nicht, oder anders formuliert, welche strukturellen Eigenschaften der Items ihre Indikatorfunktion für ein latentes Merkmal bestimmen.

Das scheinbar naheliegende Verfahren, in Anlehnung an das Kontentvaliditätskonzept zunächst Aufgabengrundmengen über klassifizierte Situations- und Verhaltenskomponenten zu definieren und anschließend empirisch zu prüfen, ob die so festgelegten Konstruktionsprinzipien geeignet sind, Aufgaben zu entwickeln, die den formalen Kriterien eines latenten Testmodells genügen, ist unter entwicklungsstrategischen Gesichtspunkten unökonomisch, wenn auch u.U. zielführend. Erweisen sich nämlich Itemteilmengen bei Anwendung eines solchen Modells als ungeeignet, könnten mittels struktureller Analysen dieser Itemmenge mögliche Ursachen mangelnder Modellverträglichkeit gefunden und die Konstruktionsregeln in einem mehrphasigen experimentellen Verfahren ggfs. so lange revidiert werden, bis Aufgaben resultieren, die mit hoher Wahrscheinlichkeit den durch das Testmodell gegebenen Anforderungen genügen. Die Modellverträglichkeit müßte jedoch nach jeder Revision der Konstruktionsregeln mit einem neuen Datensatz und einer anderen Itemstichprobe überprüft werden.

Testmodelle haben keinen unmittelbaren, inhaltlichen Bezug. Sie können daher als GEGENSTANDSUNSPEZIFISCH gekennzeichnet werden. Hypothesen über Konstruktionsregeln für potentiell testmodellkonforme Aufgaben eines bestimmten Typs sind über die Definition itemgenerierender Merkmale daher nur gewissermaßen induktiv erschließbar.

LOEVINGER (1957) schlägt schon früh vor, von einer GEGENSTANDS-SPEZIFISCHEN theoretischen Orientierung bei der Aufgabenkonstruktion auszugehen. Dieser Vorschlag steht dabei nicht unvermittelt neben dem problematisierten, in erster Linie die Außenbeziehungen der Variablen betonenden Verständnis von Konstruktvalidität, sondern ist vielmehr eingebettet in eine dreistufige Konzeptualisierung dieses Validitätsbegriffs. Unter "substantive validity" versteht LOEVINGER (1957, S.661), "the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement." Der zweite Aspekt, "structural validity", bezeichnet "the extent to which structural relations between test items parallel the structural relations of other manifestationes of the trait being measured." (S.661). Mit dem Begriff "external validity" bezieht sie schließlich das übliche Verständnis von Konstruktvalidität als Ausmaß der Beziehungen des Tests zu anderen Tests bzw. Kriterienvariablen ein.

Die beiden erstgenannten, hier vordringlich interessierenden Aspekte betonen die PARALLELISIERUNG von objektiven Aufgabenmerkmalen einerseits und den durch das jeweilige Testmodell angenommenen Verhaltenskonsistenzen andererseits. Forschungsstrategisch folgt dieser Vorgang dem angesprochenen experimentellen Testvalidierungsprozeß, indem empirische Modellkontrollen, Selektion inkonformer Aufgaben, Redefinition der Aufgabeninhalte, empirische Modellkontrollen u.s.w. logisch abhängige, aufeinanderfolgende Stufen repräsentieren. Mit Bezug auf diesen Konstruktvalidierungsvorgang bemerkt MESSICK (1975, S. 961):

> "These analyses offer evidence for the substantive component of construct validity to the extent that the resultant content of the test can be accounted for by the theory of the trait (...)".

Bedenkt man, wie Tests in der Regel validiert wurden, so muß man feststellen, daß LOEVINGER's Konzept zur Konstruktvalidierung

zumindest in der Vergangenheit kaum nennenswerte Beachtung gefunden hat, obwohl es die Dringlichkeit einer zusätzlichen inhaltlichen Orientierung bei der Testvalidierung deutlich unterstreicht. Dies mag, wie schon erwähnt, daran liegen, daß man in der psychologischen Theoriebildung lange Zeit weniger an strukturellen als an funktionalen Zusammenhängen interessiert war (vgl. WHITELY, 1982). Hinzu kommt, daß ein experimenteller Konstruktvalidierungsprozeß unter ungünstigen Bedingungen äußerst zeitaufwendig und kostenintensiv ist und damit die ökonomischen Rahmenbedingungen einer Test- oder Itembankkonstruktion in den meisten Fällen sprengen dürfte.

Weniger offensichtlich sind einige theoretische Beschränkungen, die mit dieser Konzeption der Konstruktbildung und -evaluation verbunden sind. Das Problem besteht darin, daß die empirische Gültigkeit gegenstandsunspezifischer statistischer Test- (Analyse-) Modelle nicht als explizites Bewährungskriterium für eine inhaltlich theoretische Konzeption des hypothetischen Konstrukts oder einer latenten Dimension fungieren kann. Der Interpretationsrahmen des interessierenden latenten Merkmals soll in LOEVINGER's Fassung von "substantive validity" über Aufgabeninhalte, d.h. genauer, über die durch sie bedingten psychischen Verarbeitungsprozesse bestimmt sein. Die Prüfung der Gültigkeit dieser Interpretation kann durch die gängigen Testmodelle nur implizit erfolgen, da sie gesetzmäßige Zusammenhänge zwischen den Ergebnissen dieser Verarbeitungsprozesse in Form von Testwerten/ Itembeantwortungen und des angenommenen latenten Merkmals beschreiben.

Erst die modellhafte Repräsentation präexperimenteller Annahmen über Komponenten/ Prozesse oder Strategien der Aufgabenlösung in einer Meßtheorie, die hinreichende Bedingungen für die Plausibilität eines latenten Merkmals formalisiert, gestattet letztlich "substantive validity" eines Tests EXPLIZIT über die Kontrolle aller modellkonstituierenden Annahmen für eine Personen- und

Aufgabenpopulation zu prüfen.

Es ist offensichtlich, daß die Validitätsfrage bei dieser Sichtweise nicht mehr allein über die Außenbeziehungen eines (nicht-operationalisierten) Merkmals (CRONBACH und MEEHL, 1955) beantwortet werden soll. Apriorisch angenommene, das interessierende Merkmal konstituierende Verarbeitungsprozesse, sind nun ihrerseits als Bestandteile eines zu validierenden, strukturellen Netzwerkes von Beziehungen zwischen TEILvariablen auffaßbar, weshalb KLAUER (1984, S.18) neuerdings vom "Binnenaspekt der Konstruktvalidität", EMBRETSON (WHITELY, 1980, 1982, 1983) von "Construct Representation" und SPADA und FREY (1973) sowie STERNBERG (1977) von "Interner Validierung" sprechen.

2.2.1 KONZEPTE INTERNER TESTVALIDIERUNG

Ähnlichkeiten in der Begriffswahl täuschen leicht darüber hinweg, daß o.a. Autoren in unterschiedlichen Forschungstraditionen mit ursprünglich nicht aufeinander bezogenen, ja teilweise sogar komplementären Erkenntniszielen und -möglichkeiten stehen.

Im vorliegenden Zusammenhang soll daher zunächst der Frage nachgegangen werden, welche der oben angesprochenen Konzeptionen interner Testvalidierung am ehesten geeignet scheint, die angestrebte meßzielorientierte Konstruktion von Testaufgaben theoretisch zu begründen.

2.2.1.2 STERNBERG'S "COMPONENTIAL ANALYSIS"

Einen genuin nicht-psychometrischen Ansatz verfolgt STERNBERG mit seiner "componential analysis" von Intelligenztestitems.

STERNBERG (1977) versucht, mathematische Modelle der an der Lösung gebräuchlicher Intelligenztestaufgaben beteiligten Prozesse -von ihm als "Komponenten" bezeichnet- mit kognitionspsychologischen - experimentellen Methoden zu validieren. Die Analyseeinheiten (Komponenten) faßt er als elementare Prozesse der Informationsverarbeitung auf. Definition und Verwendungsweise des Komponentenbegriffs, die sich stark an NEWELL und SIMON's (1972) Formulierung von "elementary information processes" anlehnen (vgl. STERNBERG, 1979, S.214), verdeutlichen ebenso, wie die Präferenz für Latenzzeiten als abhängige Variable seiner Untersuchungsdesigns, das seine Konzeption in der Informationsverarbeitungstheorie verankert ist.

Die Informationsverarbeitungstheorie, wie auch ihr näheres Umfeld der Kognitions- bzw. Denkpsychologie sind nun TRADITIONELL wenig

oder gar nicht daran interessiert, individuelle Merkmalsunterschiede zu beschreiben oder zu erklären. Personenvarianz wird im Gegenteil als störende "Fehlervarianz" bei der Suche nach allgemeinen Gesetzmäßigkeiten der Informationsverarbeitung bzw. von Denkabläufen aufgefaßt. Gerade aber die Nichtberücksichtigung interindividueller Unterschiede wird als hauptsächlicher Grund dafür gesehen, daß kognitions- bzw. denkpsychologische Modelle die Theorie und Praxis der Messung kognitiver Fähigkeiten kaum nachhaltig beeinflußt haben (vgl. SPADA, 1976; KLUWE und SPADA, 1981). Die Analyse von Problemlöseleistungen zugrundeliegenden Prozessen unter Ausschluß individueller Unterschiede hat zu Theorien und Modellen spezifischer, zumeist "überlernter" kognitiver Fähigkeiten geführt, die zumeist aufgabenspezifisch und nur für extrem homogenisierte Populationen gültig sind.

STERNBERG möchte die informationstheoretische Sichtweise für die Intelligenztheoriebildung nutzbar machen, indem er individuelle Unterschiede in Bearbeitungsprozessen von Aufgaben, wie sie in standardisierten psychometrischen Intelligenztests auftreten, explizit zu berücksichtigen versucht. Mit dieser Absicht folgt er der schon vor dreißig Jahren von CRONBACH (1957) erhobenen Forderung, die hauptsächlich mit korrelationsstatistischen faktorenanalytischen Methoden arbeitende Persönlichkeitsforschung und die nach allgemeinen Gesetzmäßigkeiten forschenden, Theorien entwickelnden und prüfenden experimentellen Psychologie, konzeptuell zu integrieren. Er versucht dies zunächst für den Aufgabenbereich "Analogiedenken". In seiner KOMPONENTENTHEORIE des Analogiedenkens spezifiziert er a) sechs Komponenten (incoding, inference, mapping, application, justification und response-preparation; vgl. STERNBERG, 1977) und b) eine Kombinationsregel, die die Art der Verkettung der Komponenten hypostasiert. STERNBERG nimmt an, daß sich die Lösungszeit für eine Analogieaufgabe des Typs A:B::C:D additiv aus den Responsezeiten zu jeder einzelnen Komponente zusammensetzt; jede Komponente wird durch einen Responsezeitparameter repräsentiert. Eine solche Theorie ist zu allge-

mein, um empirisch prüfbar zu sein. Sie muß durch ein Modell, das den Lösungsprozeß, oder genauer, die Art und Weise der Ausführung jeder einzelnen Komponente bei einer Aufgabe (bzw. einem Aufgabentyp) beschreibt, spezifiziert werden. Ein KomponentenMODELL beschreibt zusätzlich zur Theorie a) die Reihenfolge der Komponentenausführung und b) den Modus der Ausführung jeder einzelnen Komponente (vgl. STERNBERG, 1977, Kp. 4). STERNBERG entwickelt vier, bezüglich beider angesprochenen Aspekte konkurrierende Modelle und veranschaulicht sie anhand von Flußdiagrammen. Um die konkurrierenden Modelle empirisch prüfbar zu machen, werden mathematische Gleichungen formuliert, die die jeweilige Responsezeit als Summe der für jede involvierte Komponente benötigten Bearbeitungsdauer beschreiben. Tritt eine Komponente im Lösungsprozeß einer Aufgabe mehrmals auf, so wird der entsprechende Parameter mit seiner Auftretenshäufigkeit gewichtet.

Mit seiner Komponentenanalyse- Methode versucht STERNBERG die Angemessenheit der verschiedenen Modelle zur Beschreibung des Aufgabenlösungsprozesses abzuschätzen; das geeignetste Modell soll dann in Verbindung mit der zugrundeliegenden Theorie das Lösen von Analogieaufgaben erklären.

Eine Komponentenanalyse besteht aus einer Reihe INTENSIVER Aufgabenanalysen, deren Ergebnisse die Grundlage einer EXTENSIVEN Analyse bilden. Das Ziel der intensiven Aufgabenanalyse besteht darin, relevante Informationsverarbeitungskomponenten eines bestimmten Aufgabentyps (z.B. verbale Analogien) zu identifizieren und ein entsprechendes formales Lösungsmodell empirisch zu validieren. Ziel einer extensiven Aufgabenanalyse ist, die Verallgemeinerbarkeit der für eine Reihe verwandter Aufgabentypen aufgestellten Theorien und Modelle zu prüfen. STERNBERG (1977) unterzieht u.a. geometrische und verbale Analogieaufgaben einer intensiven Analyse, um dann in einer extensiven Analyse die Gültigkeit der Theorie- bzw. Modellelemente für den gesamten Bereich "Analogieaufgaben" aufzuzeigen.

Die intensive Analyse eines Aufgabentyps hat zwei Aspekte, den der internen und den der externen Validierung. In der Phase interner Validierung wird versucht, den gesamten Bearbeitungsprozeß bei einer Aufgabe in die zugrundeliegenden Komponenten aufzuspalten. Dabei wird die Aufgabe nach informationsverarbeitungstheoretischen Gesichtspunkten zunächst in einer Reihe von Teilaufgaben (subtasks) unterteilt, die jeweils nur eine Subgruppe oder im Idealfall genau eine, aller in der Aufgabe verwendeten Komponenten enthalten. Wenn jede, der nacheinander dargebotenen Teilaufgaben sukzessiv weniger Informationsverarbeitung erfordert, sollten die, den jeweiligen Verarbeitungsprozeß charakterisierenden, Reaktionszeitintervalle sukzessive kürzer werden. STERNBERG's allgemeine Strategie der Subtask-Analyse besteht darin, den Probanden (Pbn) nach und nach mehr lösungsrelevante Informationen zu geben, so daß für den Gesamtlösungsprozeß selbst weniger Verarbeitungszeit erforderlich ist. Die bei den einzelnen Subtasks ermittelten Responsezeit-Scores repräsentieren die Verarbeitungszeit der jeweils involvierten Subgruppe der Informationsverarbeitungskomponenten und dienen in multiplen Regressionen als Grundlage für die Schätzung der Komponentenparameter.

EMBRETSON (-WHITELY, 1983) gibt ein einfaches Beispiel, wie für verbale Analogieaufgaben Subtasks konstruiert und Annahmen über den Einfluß und das Zusammenwirken lösungsrelevanter Komponenten auf die relative Lösungshäufigkeit solcher Items mathematisch modelliert werden können. Das nachfolgende Beispiel entspricht nicht exakt dem oben beschriebenen Vorgehen STERNBERG's, da es eine modifizierte Form der Subtask-Bildung demonstriert und die Lösungsgüte und nicht, die von STERNBERG präferierten Latenzzeiten modelliert. Dennoch charakterisiert es im wesentlichen Idee und Vorgehen der Komponentenanalyse (vgl. STERNBERG, 1977, 1980, 1981; PELLEGRINO und LYON, 1979).

Abbildung 1 stellt ein Gesamtitem und zwei Subtasks dar, wie sie von WHITELY (1981) zur Analyse alternativer Modelle des Lösungsprozesses verbaler Analogieaufgaben verwendet werden.

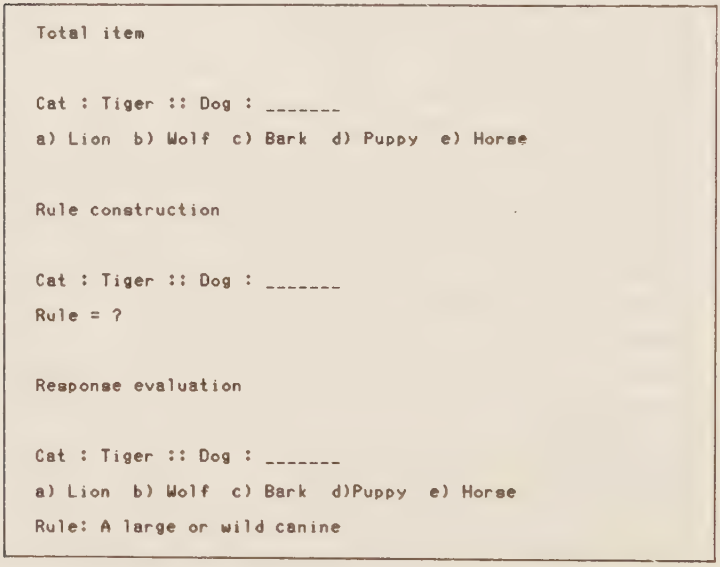


Abbildung 1:
Beispiel eines Subtask-Set für verbale Analogieitems, aus EMBRETSON (1983, S.183)

Jede Versuchsperson bearbeitete in der Untersuchung WHITELY's (1981) das Gesamtitem und einge Wochen später die einzelnen Subtasks.

Ein einfaches multiplikatives Komponentenmodell sagt die Lösungswahrscheinlichkeit für das Gesamtitem P(gT) als Produkt der Lösungswahrscheinlichkeiten P(g1) und P(g2) der beiden angenommenen Subtasks "rule construction" (g1) und "response evaluation"

(g2) voraus.

(1) $$P_{gT} = P_{g1} \; P_{g2}$$

Diese Modellformulierung verdeutlicht die Annahme eines "rationalen" Problemlösers, der z.B. nicht rät, sondern bei der Aufgabenlösung konsistent eine durch die beiden Komponenten repräsentierte, regelorientierte Strategie verfolgt. WHITELY (1977, 1981) und PARSEGHIAN et al. (1979), die ebenfalls verbale Analogien mit ähnlichen Subtasks untersuchten, berichten zufriedenstellende Prognosen der Aufgabenschwierigkeit mit Hilfe solcher Modelle.

Nur bei wenigen Aufgabentypen wird es möglich sein, Subtasks experimentell so zu isolieren, daß spezifische Effekte der experimentellen Situation (vgl. z.B. CAMPBELL und STANLEY, 1963) kontrolliert oder a priorisch ausgeschlossen werden können.

Liegen, wie dies üblicherweise der Fall ist, Daten über richtige und falsche Aufgabenbearbeitungen vor, so können über angenommene "Komplexitätsfaktoren" Hypothesen über relevante Struktur- und Prozeßvariablen der Itembearbeitung entwickelt und geprüft werden. Der Grundgedanke besteht darin, die kognitive Komplexität einer Aufgabe, die ihre Schwierigkeit ausmacht, als Funktion der Anzahl und der Schwierigkeit der zur Aufgabenlösung erforderlichen kognitiven Operationen oder Prozesse darzustellen (vgl. BOURNE, EKSTRAND und DOMINOVSKI zur Methode des "Hypothesis Testing" in der Denkpsychologie).

Um die Methode der "Komplexitätsfaktoren" wieder an einem einfachen Beispiel zu demonstrieren, sei hier kurz auf die Untersuchung geometrischer Analogieaufgaben von MULHOLLAND, PELLEGRINO und GLASER (1980) hingewiesen.

Um eine Aufgabe der Art, wie sie in Abbildung 2 dargestellt ist,

zu lösen, sind, so die Hypothese der Autoren, zwei unabhängige Verarbeitungsprozesse erforderlich: a) "encoding complexity", abhängig von der Anzahl der Elemente in Stimulus A und b) "transformational complexity", die von der Anzahl der Operationen abhängen soll, die erforderlich sind, um Stimulus A in Stimulus B zu transformieren.

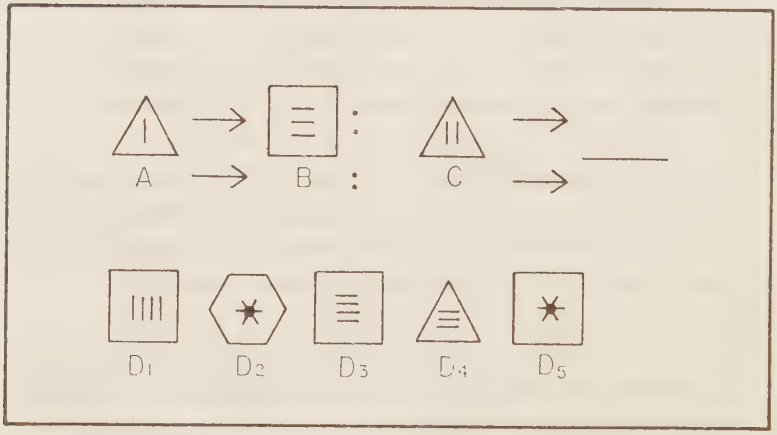


Abbildung 2 :
Regelgeleitet konstruierte geometrische Analogieaufgabe (aus EMBRETSON, 1982)

In der abgebildeten Aufgabe besteht Stimulus A aus zwei Elementen (Dreieck, Strich) und drei Operationen sind erforderlich, um Stimulus A in B zu transformieren (Veränderung des äußeren Figurelements, Erhöhung der Anzahl innerer Figurelemente und deren Rotation um ein 90 Grad-Winkel).

WHITELY und SCHNEIDER (1981) nehmen weiter an, daß sich die Art der Transformation ("spatial displacement" versus "distortions",

S.385) differentiell auf die Itemschwierigkeit auswirkt. Um das Beispiel nicht zu komplizieren, sollen konkurrierende Hypothesen zunächst außer Acht gelassen werden.

In einem einfachen Modell läßt sich die Schwierigkeit einer geometrischen Analogieaufgabe als gewichtete Summe der beiden Komplexitätsfaktoren "encoding" und "transformations" darstellen:

$$(2) \qquad P_g = f_{1g} n_1 + f_{2g} n_2 + c$$

wobei P(g) die Schwierigkeit der Aufgabe g, f(mg) die Häufigkeit des Prozesses m in Item g und n(m) seinen Beitrag zur Schwierigkeit der Aufgabe g bezeichnen.

MULHOLLAND et al. (1980) fanden in ihren Untersuchungen Evidenzen für die Gültigkeit des 2-Komponentenmodells, wobei sie jedoch exponentielle Beziehungen zwischen Itemmittelwerten/ mittleren Latenzzeiten und den Komplexitätsfaktoren annehmen.

Ein letztes Beispiel für eine Komponentenanalyse sensu STERNBERG verbindet die Subtask-Methode mit der Methode der Komplexitätsfaktoren. STERNBERG und TURNER (1981) untersuchten Aufgaben zum syllogistischen Schlußfolgern folgender Art:

Alle A sind B; alle B sind C.
a) Alle A sind C, b) Einige A sind C, c) Keine A sind C,
d) Einige A sind keine C, e) Keine der Alternativen trifft zu.

Hier wurden Subtasks konstruiert, um die beiden Verarbeitungsprozesse "Prämissenenkodierung" und "Kombination der Information über die Beziehung zwischen A und C" experimentell zu separieren. Ein EnkodierFEHLER würde beispielsweise vorliegen, wenn ein Proband davon ausgeht, daß die Prämisse "Alle A sind B" impliziert, daß "Alle B sind A". Ein typischer Kombinationsfehler bestünde darin, A über Informationen bezüglich B mit C in Beziehung zu

setzen.

STERNBERG und TURNER (1981) postulieren als zusätzliche relevante Komponente der Aufgabenlösung einige Komplexitätsfaktoren, mit denen sie die Art der Prämissen zu beschreiben versuchen (z.B.: partikuläre versus universelle, affirmative versus negative Prämissen). Den Autoren gelang es, ein Modell des syllogistischen Schließens auf der Grundlage von Subtask-Beantwortungen und der Komplexitätsfaktoren zu validieren. Dies zeigt, daß sich beide Methoden der Komponentenanalyse keineswegs ausschließen.

Es ist unbestritten, daß die hier nur global umrissene Komponentenanalyse im Sinne STERNBERG's einen Beitrag zur internen Validierung von Intelligenztests, d.h. zum Verständnis der erfaßten latenten Merkmale/ Fähigkeiten leistet. RUDINGER und RÜPPEL (1981) sehen in STERNBERG's Ansatz sogar einen ersten Schritt, die Probleme traditionell entwickelter Intelligenztests zu überwinden und zu eindeutigen und bedeutsamen Quantifizierungen zu gelangen:

> "STERNBERG's quantitativ - qualitatives Vorgehen ist u.E. der augenblicklich komplexeste und durchdachteste Ansatz der Intelligenzforschung, der die psychometrische Tradition mit derjenigen der Informationsverarbeitung verbindet. Wenn der STERNBERG'sche Ansatz oder ähnliche (vgl. RESNICK, 1976) erfolgreich sein sollten und sich ein Satz stabiler Komponenten ermitteln ließe, der empirisch abgesichert wäre, dann wäre unter meßtheoretischem Aspekt das (...) Problem der Meßbarkeit der Intelligenz sicherlich einer Lösung näher gebracht" (S.213).

Im vorliegenden Zusammenhang interessiert in erster Linie, ob aus empirisch erprobten mathematischen Komponentenmodellen Aussagen darüber ableitbar sind, wie Aufgaben konstruiert werden können, die konsistente, interindividuelle Unterschiede bei einem, über psychologische Verarbeitungsprozesse definierten, latenten Merk-

mal messen.
Folgt man zunächst der zuversichtlichen Einschätzung von RUDINGER und RUPPELL (1981), so sollte dieses Ziel wohl erreichbar sein, wenn sich ein "stabiler", was eigentlich nur heißen kann, über einer Item- und Personenpopulation verallgemeinerbarer, Satz von "Komponenten" finden und empirisch validieren ließe.

Dies bedeutete beispielsweise für den von STERNBERG (1977) intensiv analysierten Typus der Analogieaufgaben, daß die Lösungsgeschwindigkeit und/oder die Lösungszeiten bei UNTERSCHIEDLICH schwierigen Analogieitems als eine durch das mathematische Modell spezifizierte Funktion der Art und Zahl INVARIANTER, d.h. von der untersuchten Personenstichprobe unabhängiger "performance"-Komponenten dargestellt werden können.

Zentrale Annahmen der Komponentenmodelle STERNBERG's lassen jedoch, ebenso wie die experimentellen Randbedingungen ihrer Validierung, schwerwiegende Zweifel an ihrer Bedeutsamkeit als allgemein gültige, und damit insbesondere für die Itemkonstruktion brauchbare, Aufgabenbearbeitungstheorien aufkommen (vgl. KIRCHENKAMP, 1984). Das prinzipielle Dilemma in STERNBERG's Ansatz resultiert aus den Versuchen, Modelle der Lösungszeit und der Lösungsgüte SIMULTAN, d.h. unter gleichen experimentellen Bedingungen, zu validieren. Um sinnvolle Aussagen über relative oder absolute Beiträge hypostasierter Elementarprozesse an der GesamtbearbeitungsDAUER einer Aufgabe machen zu können, ist STERNBERG auf Meßdaten angewiesen, die er als Ausführungszeiten einzelner Komponenten interpretieren kann. Solche Daten können natürlich nur dann vorliegen, wenn die fraglichen Komponenten tatsächlich ausgeführt werden und damit zu korrekten Beantwortungen der Subtasks und letztlich der Gesamtaufgabe führen. Latenzzeiten aus Falschbeantwortungen verfälschen die Parameterschätzungen einzelner Komponenten oder lassen sie völlig uninterpretierbar werden. Die Validierung solcher Latenzzeitintervalle

ist daher nur unter experimentellen Bedinungen möglich, die sich von typischen Testvorgabesituationen im wesentlichen dadurch unterscheiden, daß die Personenvarianz bewußt minimiert wird. Dies kann entweder dadurch geschehen, daß einer unausgewählten oder untrainierten Personenstichprobe ausschließlich solche Testaufgaben vorgelegt werden, die selbst in solch heterogenen Stichproben extrem hohe Lösungshäufigkeit aufweisen, oder indem eine "hochbefähigte" Personenstichprobe in der Bearbeitung schwierigkeitsheterogener Aufgaben trainiert wird.

STERNBERG konstruiert diese "künstliche" Situation, indem er vorhandene Intelligenztestaufgaben teilweise vereinfacht und durch ein gezieltes Training die geforderten Problemlösungskomponenten an ähnlichen Aufgaben überlernen läßt.

Durch die experimentelle Veränderung der Test- bzw. Aufgabenvorgabebedingungen wird die Verallgemeinerbarkeit der gewonnenen Daten und Ergebnisse, also die externe Gültigkeit des Experiments (vgl. COOK und CAMPBELL, 1979) fraglich; beim Überlernen des Aufgabenmaterials verändert sich mit einiger Wahrscheinlichkeit der kognitive Anforderungsgehalt, d.h. Art und Zahl zur Aufgabenlösung vermutlich erforderlichen Operationen und Prozesse. Es ist anzunehmen, daß die von STERNBERG postulierten Theorien und Modelle allenfalls "ein algorithmisches Vorgehen für die Lösung sehr einfacher Items oder automatisierte Prozesse" (KIRCHENKAMP, 1984, S.82) abbilden.

STERNBERG (1977) berichtet relative Aufgabenlösungswahrscheinlichkeiten, die zwischen p=.85 und p=.95 liegen. Mit dem gleichen Komponentenmodell, mit dem es STERNBERG gelingt, bezogen auf Gruppendaten, immerhin zwischen 62% und 93% der Varianz mittlerer Latenzzeiten aufzuklären, konnten lediglich 12% bis 59% der -geringen- Fehlerdatenvarianz erklärt werden.

Die nur schwache bis mittlere Erklärungskraft der angenommenen Komponenten im Sinne schwierigkeitserzeugender Faktoren der Aufgabenlösung ist nicht der einzige Grund, weshalb dieser Ansatz interner Konstruktvalidierung für die Belange einer theoretisch angeleiteten Aufgabenkonstruktion bezüglich einer Leistungsgütedimension kaum brauchbar scheint. Neben der schon angesprochenen Spezifität der Bearbeitungsprozesse bei überlerntem Itemmaterial, sprechen weitere inhaltliche, wie auch formale Beschränkungen des Ansatzes gegen seine Verwendung bei der Konstruktvalidierung und Testkonstruktion.

Der erste Einwand betrifft die Quantifizierung der Itemschwierigkeit bzw. Bearbeitungszeit und der involvierten Operationen/ Prozesse in mathematischen Modellen der Komponentenanalyse. Die modellierten mittleren Aufgabenbeantwortungszeiten und relativen Lösungshäufigkeiten sind in starkem Maße von der Zusammensetzung der untersuchten Personenstichproben abhängig. In, bezüglich des fraglichen Merkmals oder relevanter Kovariate, homogen(isiert)en Untersuchungsstichproben haben diese Indizes eine geringere Variabilität als in repräsentativeren Populationen. Folglich werden auch Komplexitätsfaktoren repräsentierende Regressionsgewichte (vgl. Gleichung 2) mitunter nicht mit denen aus heterogenen Stichproben übereinstimmen. BezugsgruppenINVARIANTE Quantifizierungen komplexitätserzeugender Operationen und Prozesse sind aber insbesondere dann notwendig, wenn aus der Aufgabenbeantwortungstheorie Prognosen über erwartete Schwierigkeiten/ Bearbeitungszeiten zu konstruierender Aufgaben für bestimmte diagnostische Problemstellungen abgeleitet werden sollen. Die nächsten Kapitel diskutieren spezifische Voraussetzungen und Möglichkeiten einer in diesem Sinne theoretisch angeleiteten Aufgabenkonstruktion.

Quantifizierungen der Lösungskomponenten- und Itemschwierigkeiten mit stichprobenabhängigen Indizes beschränken den Gültigkeitsbe-

reich von Komponentenmodellen auf die analysierten Aufgaben und die rekrutierten Versuchspersonen. Sie sind aber selbst als "Miniaturtheorien" (SPADA, 1976, S.157) unbrauchbar, wenn die angenommenen Komponenten nicht eindeutig interpretiert werden können, weil die Schätzung ihrer Schwierigkeit in unbekanntem Maße von den übrigen Faktoren der Aufgabenlösung abhängt. Die Gefahr einer "expliziten Parameterkonfundierung" (PELLEGRINO und LYON, 1979) ist besonders dann gegeben, wenn bestimmte Komponenten in experimentellen Subtasks stets gemeinsam auftreten. So ist häufig nicht entscheidbar, ob sich Komponenten wiederholen oder verschiedene Komponenten ausgeführt werden.

> "If the component scores are very highly correlated across subjects (after correction for attenuation) then there is a good chance that the component scores are actually measures of the same component. One cannot be certain, however, since distinct components may be correlated" (STERNBERG, 1977, S.76).

PELLEGRINO und LYON (1979) warnen außerdem vor "impliziten Konfundierungen", die immer dann vorliegen, wenn Komponentenparameterschätzungen zusätzliche, theoretisch bzw. modellhaft nicht abgebildete Prozesse widerspiegeln. Nach Ansicht der Autoren sind implizite Konfundierungen im Datenmaterial STERNBERG's wahrscheinlich (vgl. dazu auch KIRCHENKAMP, 1984).

Bei eingeschränkter interner Validität der Komponentenmodelle ist nicht zu erwarten, daß in "extensiven Analysen" der Nachweis allgemeiner, aufgabentypübergreifender Komponenten gelingt. Entsprechende Generalisierungsversuche für verschiedene Analogieaufgabentypen ("people-piece", "verbal", "geometric") zeigen denn auch, daß die Komponentenscores über Itemtypen hinweg erheblich stärker variieren als innerhalb eines bestimmten Typs (vgl. STERNBERG, 1977; SIMON, 1979). Eine alternative Möglichkeit, die Allgemeinheit der Komponenten zu demonstrieren, bestünde ferner darin, sie als materialunabhängige Quelle konsistenter, indivi-

dueller Unterschiede auszuweisen. Schätzungen der Personenparameter aus Individualdaten erwiesen sich bei GARTNER und STERNBERG (1983) jedoch als derart unreliabel, daß auf diese Validierungsmöglichkeit verzichtet werden mußte:

> "Unfortunately, the latency parameters for individual subjects were of insufficient reliability to allow tests of individual differences predictions at the component level" (S. 15).

Daß die Komponentenmodelle STERNBERGS's die Komplexität der untersuchten Aufgaben für den avisierten Zweck der theoretisch angeleiteten Aufgabenkonstruktionen nur unzureichend zu erklären vermögen und individuelle Unterschiede in der Fähigkeit die angenommenen Komponeten auszuführen, nicht reliabel beschrieben werden können, ist jedoch möglicherweise weniger unangemessenen Hypothesen über den Aufgabenlösungsprozeß, als vielmehr einer grundsätzlich inadäquaten Modellbildung und Parametrisierung zuzuschreiben. STERNBERG strebt einerseits möglichst allgemeingültige Aussagen über strukturelle Elemente seiner Aufgabentheorien (Komponentenschwierigkeiten etc.) an und möchte andererseits individuelle Unterschiede bezüglich dieser Strukturelemente beschreiben. Dabei übersieht er, daß beide Problemstellungen nicht unabhängig voneinander lösbar sind, denn auch die allgemein(-psychologischen) "Anteile" seiner Komponententheorien können nur über Beobachtungen an Individuen validiert werden. Jede Einzelbeobachtung muß aber als das Ergebnis einer (möglicherweise) allgemeingültigen Gesetzmäßigkeit UND differentieller Verhaltenstendenzen betrachtet werden. Im vorliegenden Fall "partiell konsistenter Beobachtungen" (NEYMAN und SCOTT, 1948, zit. nach FISCHER, 1971) werden für eine angemessene Beschreibung der Daten gleichzeitig allgemeine, "strukturelle" und individuelle, oder in der Diktion von NEYMAN und SCOTT (1948), "inzidentelle" Parameter benötigt (vgl. SCHEIBLECHNER, 1972). FISCHER (1971) verdeutlicht den Unterschied zwischen den

Parameterarten am Beispiel der Faktorenanalyse:

> "Die Faktorenladungen sind Strukturparameter (...), weil sie sich auf Struktureigenschaften der Intelligenz, also auf die STRUKTUR DER THEORIE beziehen. Faktorenwerte sind inzidentelle Parameter, weil sie von den ZUFÄLLIG in die Stichprobe rekrutierten Vpn abhängen (...). Der Unterschied zwischen Strukturparametern und inzidentellen Parametern liegt nur in ihrer Anzahl. Die Strukturparameter sind von endlicher Anzahl und kommen in unendlich vielen beobachtbaren Zufallsvariablen vor (...). Demnach ist es möglich über jeden Strukturparameter beliebig viel Information zu sammeln, d.h. ihn beliebig genau zu identifizieren. Im Gegensatz dazu gibt es unendlich viele inzidentelle Parameter, von denen jeder einzelne in höchstens endlich vielen Beobachtungen enthalten ist. (...) Also können inzidentelle Parameter nicht beliebig genau erfaßt werden (S. 379)."

Nach einem Satz von NEYMAN und SCOTT (1948) können die für die Theoriebildung interessanten "Strukturparameter" nicht in befriedigender Weise geschätzt werden, wenn ihre Schätzung, wie etwa in der Faktorenanalyse, zugleich mit der Schätzung inzidenteller Parameter erfolgen muß. Aufgabenschwierigkeiten und die sie bedingenden Faktoren werden in einer formalen Itembeantwortungstheorie durch strukturelle Parameter beschrieben. Ihre theoretische Brauchbarkeit für eine prognostisch orientierte Aufgabenkonstruktion hängt also nicht allein von der Adäquanz inhaltlicher Theorien, sondern auch davon ab, wie weit es gelingt a) gleichzeitig beide Parameterarten zu berücksichtigen und b) die Parameter unabhängig voneinander zu schätzen.

Beide Bedingungen werden von den Komponentenmodellen STERNBERG's und anderer Autoren, die sich dieser Art mathematischer Modellbildung bedienen, nicht erfüllt. Auf die negativen Folgen der Stichprobenabhängigkeit von Komponentenparameterschätzungen für ihre Verallgemeinerbarkeit wurde schon hingewiesen. Die fehlende Berücksichtigung von Personenparametern in den strukturellen Modellen führt außerdem dazu, daß diese Modelle keinerlei Beziehungen zwischen den Eigenschaften der Aufgaben und dem Antwortverhalten der Pbn spezifizieren. Unabhängig davon, ob nun Itemmittelwerte oder Individualdaten als abhängige Variable in die Regressionsgleichungen eingehen, der Gegenstand betreffender Modelle ist stets die Aufgabenvarianz und nicht die Personenvarianz. Es besteht demnach keine Möglichkeit zu antizipieren, wie die Relation zwischen diesen beiden Varianzquellen aussieht. So ist beispielsweise ein Test mit nahezu uniformer Aufgabenschwierigkeit denkbar, der zuwenig reliable Itemvarianz aufweist, um ein mathematisches Strukturmodell sinnvoll anpassen zu können. Die Testdaten weisen hingegen eine hohe Personenvarianz auf, wenn man weiter annimmt, daß die relativen Aufgabenlösungshäufigkeiten nahe p=.50 liegen. Der Extremfall verdeutlicht, daß ein solches Strukturmodell die Quellen der PERSONENvarianz nicht eindeutig in den angenommenen Komponenten der Aufgabenlösung zu lokalisieren vermag. Unterscheiden sich Personen, wie im obigen Beispiel, in ihrer Antwortvarianz oder Vorhersagbarkeit der Antworten aus dem mathematischen Modell, dann haben die Regressionsgewichte keine vergleichbare Metrik über Personen hinweg. Bei den gängigen binären Antwortdaten hängt die Varianz der Itembeantwortungen direkt von der Fähigkeit der Person zur Lösung solcher Aufgaben ab und je mehr Personen in ihrer Fähigkeit variieren, desto unangemessener werden die Komponentenparameterschätzungen für die Messung interindividueller Fähigkeitsunterschiede sein.

2.2.1.2 "CONSTRUCT REPRESENTATION" SENSU EMBRETSON

Die vielfältigen Beschränkungen der Komponentenanalyse verdeutlichen, daß die Erfolgsaussichten, eine für die meßzielorientierte Itemkonstruktion brauchbare Aufgabenbeantwortungstheorie zu entwickeln, nicht allein von angemessenen Hypothesen über den Lösungsprozeß, sondern insbesondere auch von einer adäquaten PSYCHOMETRISCHEN Modellbildung und -prüfung abhängt.

EMBRETSON (-WHITELY, 1980, 1982, 1983) setzt mit ihrem Entwurf der "construct representation" genau bei dieser Argumentation auf und stellt vier allgemeine Forderungen an eine effektive Methodologie der internen Konstruktvalidierung:

"First, (...) the relation of person performance to item characteristics must be clearly specified in the methodology.

Second, the methodology must provide for the comparison of alternative theories of the constructs that are involved in the task. (...)

Third, the methodology must provide for quantification of the theoretical constructs in specified items. (...)

Forth, the methodology should provide for the measurement of persons on the constructs." (EMBRETSON, 1983, S.181).

Die ersten drei Bedingungen sind von ESSENTIELLER Bedeutung für eine konstruktionstechnologisch brauchbare Aufgabenbeantwortungstheorie. Sie sind nur schwer isoliert herstellbar, wie die Kritik an der Komponentenanalyse gezeigt haben dürfte. Sie setzen sich z.T. wechselseitig voraus, weshalb EMBRETSON (1983) durch die Rangordnung wohl keinerlei Gewichtung impliziert wissen möchte.

Erweist sich eine Aufgabenbeantwortungstheorie konkurrierenden Theorien (empirisch) überlegen (Bedingung (2)) und erfüllt sie zudem die Bedingungen (1), (2) und ggfs. (4) -in einer noch weiter zu spezifizierenden Weise-, dann ist eine TECHNOLOGIE für die meßzielorientierte, auf spezifische diagnostische Fragestellungen ausgerichtete, Konstruktion von Aufgaben eines definierten Merkmalsbereichs ableitbar.

Ist das diagnostische Interesse an dem definierten Konstrukt formuliert, oder genauer, ist aufgrund inhaltlicher - oder entscheidungstheoretischer Vorüberlegungen spezifiziert, welche kognitiven Anforderungen, repräsentiert durch Art und Komplexität der in Aufgabenlösungsprozessen involvierten Operationen/ Prozesse, mit welcher Präzision gemessen werden soll, dann ist zunächst über die in "Bedingung 1" angesprochene formale "Spezifikation" der Beziehung zwischen "person performance" und "item characteristics" angebbar, wie diese, das Item als Ganzes charakterisierenden Merkmale, beschaffen sein müssen, damit es für die Erreichung des Meßziels maximal effizient bzw. informativ ist. Beschreibt die betreffende Theorie Unterschiede zwischen Aufgaben formal über ihre verschiedenen Lokalisationen auf einer gemeinsamen Schwierigkeitsskala, so besteht das Konstruktionsziel darin, homogene/ eindimensionale Aufgaben theoretisch vorhergesagter Schwierigkeit zu entwickeln. Wenn eine valide Aufgabenbeantwortungstheorie die Aufgabencharakteristika durch quantitativ parametrisierte Lösungsoperationen/ Prozesse ERKLÄRT (vgl. Bedingung (3)), ist das jeweilige Konstruktionsziel erreichbar, indem einzelne Operationen bzw. Prozesse "bekannter Komplexität" absichtsvoll zu Aufgaben mit abschätzbaren Eigenschaften synthetisiert werden. Das schließlich gemessene Merkmal ist dann interpretierbar als eindimensionale Fähigkeit oder Disposition, ALLEN Personen einer Population prinzipiell bekannte und verfügbare Operationen/ Prozesse in einer vorgegebenen Situation (Item) "bereitzustellen" und effizient "auszuführen" (Bedingung (1) gegeben, Bedingung (4) jedoch nicht; man beachte, daß Bedingung

(4) als "Soll-" und nicht als "Muß-" Kriterium definiert ist). Ist Bedingung (4) ebenfalls erfüllt, d.h. werden mehrere, auf die angenommenen Operationen/ Prozesse bezogene Durchführungs-/ Lösungsfähigkeiten in einem mehrdimensionalem Bearbeitungsmodell getrennt parametrisiert, sind zusätzlich Aussagen über die individuelle "Operations-" oder "Prozeßausstattung" eines Pb möglich.

Welche der beiden, auch qualitativ verschiedenen Annahmen für einen gegebenen Merkmalsbereich und eine Population plausibel ist, wird im Einzelfall sowohl auf der Grundlage inhaltlich-theoretischer Überlegungen, wie aber auch empirisch zu prüfen sein.

EMBRETSON (1982) sieht den wesentlichen Gewinn einer solchen Konstruktrepräsentation ebenfalls darin, auftragsbezogene, diagnostische Zielvorstellungen in der Phase des "test design" adäquat zu berücksichtigen:

> "(...) construct representation research is designed to elaborate the theory of the item task. This theory can be used as item specifications to develop tests to measure certain factors. However, unlike traditional item specifications, this theory is explicitly tested and studied." (S.15)

Ist das Konstruktionsrational in allen Bestandteilen hinreichend spezifisch, so ist, in Konsequenz, eine (voll-)automatisierte Aufgabenkonstruktion denkbar. BEJAR (1984) betrachtet dies schon als Nahziel des "test design" - Ansatzes:

> "(...) if test developers are able to account for differences among items that may have captured the knowledge necessary to synthesize items of known characteristics (...). They may, in short, be able to write a computer programm capable of composing an item with known psychometric characteristics" (S.19).

Im Rahmen computerunterstützter Teststrategien (CAT) wären dann nicht allein die adaptive Auswahl jeweils geeigneter Aufgaben aus einem residenten Itempool, sondern auch gleichermaßen deren "adaptive" Konstruktion in jeder Phase des Testprozesses realisierbar (vgl. HABON, 1981). Die Möglichkeiten des CAT und auch des computerunterstützten Unterrichts wären dann nicht mehr unmittelbar auf die Eigenschaften des zur Verfügung stehenden Aufgabenpools limitiert, sondern könnten durch die rationale, zielorientierte Konstruktion von Test- und ggfs. Übungsaufgaben in vielfacher Hinsicht optimiert werden.

Im Gegensatz zum oben skizzierten mehrphasigen Testkonstruktions- und Konstruktvalidierungsprozeß (vgl. Kap. 2.2), basiert die rationale Aufgabenkonstruktion explizit auf der zugrunde liegenden Aufgabenbeantwortungstheorie. Jede rationale Konstruktion ist folglich ein EXPERIMENT, das die Gültigkeit des zugrundeliegenden Konstrukts prinzipiell in Frage stellt. Das Konstruktionsexperiment ist als gescheitert zu betrachten, wenn die Antwortdaten durch das Antwortmodell nicht beschrieben werden können oder, wie dies üblicherweise bei Item- und Testanalysen der Fall ist, Aufgaben selegiert werden müssen, um die Gültigkeitsbehauptung für ein existierendes Modell aufrecht erhalten zu können. Daß Aufgabenkonstruktionen auf diese Weise scheitern, ist gerade zu Beginn einer solchen Theoriebildung wahrscheinlich. Traditionelle, andernorts schon als "irrational" gekennzeichnete Testkonstruktionen (s. HORNKE, 1984), kalkulieren den "drop-out" ungeeigneter Aufgaben ein, indem von vornherein wesentlich mehr Items erstellt werden, als letztlich in der Testendform verwendbar sind. Nachfolgende Testkonstruktionen können von dieser Praxis nicht profitieren, da sie ohne direkten theoretischen Bezug, zu keinerlei Verbesserung des Konstruktions- "know-how" führt. Scheitert hingegen eine rationale Testkonstruktion, so werden die Ursachen des Scheiterns in der Regel im Konstruktionsrational lokalisierbar sein. Die Identifikation "problematischer" Konstruktionselemente, ihre Beschreibung in der Begrifflichkeit der Konstruk-

tionstheorie, wie auch ihre anschließende Revision und erneute empirische Überprüfung sind konstruktive Bestandteile des Theoriebildungsprozesses, an dessen (vorläufigem) Ende ein erklärungs- und prognosekräftiger, d.h. auch "Konstruktionsirrtümer/ -fehler" minimierender Itemkonstruktionsansatz stehen dürfte.

Rationale Aufgabenkonstruktion könnte auf diese Weise nicht nur einen wesentlichen Beitrag zur Theoriebildung leisten, sondern auch dazu beitragen, die Entwicklung großer Aufgabenpools, sogenannter Itembanken, wirtschaftlich zu gestalten (vgl. NEWBOULD und MASSEY, 1977). BEJAR (1984) betrachtet die "Rationalisierung" der Aufgabenkonstruktionen ebenfalls unter diesem ökonomischen Gesichtspunkt:

> "(...) adopting that effort as a goal puts test developers in the enviable position of simultaneously persuing scientific and economic goals. That is, the ability to synthesize items is likely to improve the productivity of the test designer in much the same way that computers have altered the productivity of, for example, graphics designers in various industries" (S.19).

Es ist zu erwarten, daß die Ablösung der nach dem "trial and error" - Prinzip verfahrenden traditionellen Aufgaben- und Testkonstruktionen durch eine den "Ausschuß" unbrauchbarer Items minimierende rationale Konstruktion kommerzielle und nichtkommerzielle Bereiche der Testdiagnostik quantitativ beleben und qualitativ verbessern wird (vgl. MILLMAN und ARTER, 1984).

Die von EMBRETSON (1983) erhobenen Forderungen an eine Methodologie der Konstruktrepräsentationen sind allgemein formuliert und bedürfen der Präzisierung im Hinblick auf die Entwicklung formaler Modelle, die allen drei bzw. vier zitierten Kriterien genügen.

3. ANFORDERUNGEN AN FORMALE MODELLE ZUR REPÄSENTATION LATENTER VERHALTENSDISPOSITIONEN

In der Testtheorie wird ein psychologischer Test definiert als "wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung" (LIENERT, 1969/3, S. 7). Eine grundlegende, auch in den zitierten "Kriterien" EMBRETSON's (1983) implizit vorausgesetzte Forderung an jedes Testmodell besteht demnach darin, daß es die zentralen Fragen der EXISTENZ und MESSBARKEIT eines oder mehrerer "Persönlichkeitsmerkmale" einer empirischen Überprüfung zugänglich macht. Dieses Problem ist formal identisch mit dem von LORD und NOVICK (1968) geforderten empirischen Nachweis einer gesetzmäßigen Beziehung zwischen einer angenommenen latenten Variable und dem manifesten, beobachtbaren Testverhalten.

Es wurde schon erwähnt (vgl. Kap. 2.2), daß dieser, auch als "Repräsentationsproblem" gekennzeichnete Sachverhalt (s. SUPPES und ZINNES, 1963) nicht von allen gängigen Testtheorien in befriedigender Weise gelöst wird. Die "klassische Testtheorie" (GULLIKSEN, 1950; LIENERT, 1969/3) und ihre Verallgemeinerung zur Faktorenanalyse (PEARSON, 1901; SPEARMAN, 1904; HOTELLING, 1933; THURSTONE, 1947; REVENSTORF, 1976) galten lange Zeit als geeignete Methode latente Dimensionen aufzuzeigen und zu beschreiben. Die Legitimation einer solchen Variable schien dann gegeben, wenn eine Menge von Items im Sinne der "simple structure" nach THURSTONE (1954), hohe Ladungen auf nur einem Faktor aufwies, andere Items hingegen auf anderen Faktoren. Vermochte man mit möglichst wenigen Faktoren die Varianz möglichst vieler Testvariablen zu beschreiben, so galt es als evident, jene Faktoren gefunden zu haben, welche als grundlegende verhaltenserklärende Persönlichkeitsmerkmale (Eigenschaften, Dimensionen, Dispositionen oder Konstrukte) interpretiert werden konnten.

Es darf nicht übersehen werden, daß die Ergebnisse früher faktorenanalytischer Untersuchungen -in der Begrifflichkeit JÖRESKOG's (1969)- eher explanatorischen als konfirmatorischen Charakters waren. Die als Ausgangsdaten dienenden Korrelationsmatrizen sind nämlich in hohem Maße von Varianzen und Kovarianzen der Testitems in der aktuell untersuchten Personenstichprobe abhängig (s. z.B. FISCHER, 1974, S.136), was zur Folge hat, daß erhebliche Artefakte auftreten können, z.B. wenn bei unterschiedlichen Stichproben für die gleichen Testitems durchaus verschiedene Faktorenlösungen auftreten. Als bemerkenswertes Beispiel sei hier nur die Differenzierungshypothese der Intelligenz erwähnt (REINERT, BALTES, SCHMIDT, 1965), deren Ergebnisse von MERZ und KALVERAM (1965) als mögliches Artefakt in Frage gestellt werden konnten (vgl. auch SPADA, 1969). Weitere Probleme, die die Rationalität faktorenanalytischer Verfahren und die Interpretation ihrer Ergebnisse betreffen, werden z.B. von FISCHER (1974), CARROLL (1978), und HABON (1981) diskutiert.

Fortschritte bahnten sich in den 50er Jahren an, als einer anderen Denkweise der Vorzug gegeben wurde. Latente Eigenschaften sollten nicht länger allein deshalb angenommen werden, weil sie (wie z.B. in Form des true-scores in der KTT oder eines Faktors in der Faktorenanalyse) möglichst viel Varianz mit ihren beobachteten Manifestationen gemeinsam haben. Vielmehr sollte versucht werden, bestimmte Eigenschaften/ Merkmale dadurch zu legitimieren, daß sie als URSACHE für die Kovarianz zwischen den Testvariablen angesehen werden können. Ein latentes, nicht direkt beobachtetes, verhaltensbestimmendes Merkmal gilt in diesem Sinne als nachgewiesen, wenn gezeigt werden kann, daß die ursprünglich beobachtbaren Korrelationen zwischen den Testvariablen verschwinden, sofern man sie an Personen mit konstanter Ausprägung in dem zu erschließenden latenten Merkmal untersucht.

Die formale Struktur dieser Schlußweise ist keineswegs neu. Sie läßt sich bereits Mitte des vorigen Jahrhunderts bei JOHN STUART

MILL auffinden. Er schreibt in seinem "Zweiten Kanon" der wissenschaftlichen Forschung:

> "Wenn eine Instanz, in der das zu erforschende Phänomen eintritt, und eine Instanz, in der es nicht eintritt, jeden Umstand bis auf einen gemeinsam haben, indem dieser eine nur in der ersten vorhanden ist, so ist der Umstand, in dem die beiden Instanzen voneinander abweichen, die (...) Ursache oder ein unentbehrlicher Theil der Ursache des Phänomens" (MILL, 1885, S.91f.).

Im vorliegenden Fall ist das "zu erforschende Phänomen" die Korrelation zwischen Testvariablen, und als Ursache gesucht ist ein latentes Merkmal, das so beschaffen ist, daß die Korrelation "eintritt", wenn das Merkmal variiert, aber "nicht eintritt", wenn das Merkmal konstant gehalten wird.

In Anlehnung an diese Heuristik wurden ab 1950 verschiedene "latente Dimensionsmodelle" publiziert (LAZARSFELD, LATENTE STRUKTURANALYSE, 1950; GUTTMAN, SCALOGRAMMANALYSE, 1950; RASCH, SPEZIFISCHE OBJEKTIVITÄT, 1960), die detailierte, prüfbare Voraussetzungen über das Reaktionsverhalten der Probanden sowie MESSTHEORETISCHE Überlegungen einbeziehen und auf diese Weise latente Verhaltensdispositionen in grundsätzlich falsifizierbarer Weise eindeutig erschließbar machen.

Die allgemeine MESSTHEORIE nach SUPPES und ZINNES (1963; vgl. etwa die Darstellung bei ORTH, 1974 oder TACK, 1970) nennt zwei Grundbedingungen, die erfüllt sein müssen, damit aus ihr bestimmte Messmodelle abgeleitet werden können:

1. Es muß die Existenz eines EMPIRISCHEN RELATIVS vorausgesetzt werden, das aus einer Menge empirisch beobachtbarer Objekte (den Merkmalsträgern) und mindestens einer auf diesen Objekten definierten empirischen Relation besteht.

2. Es muß ein NUMERISCHES RELATIV angebbar sein, das aus einer Menge reeller Zahlen und darauf definierten numerischen Relationen besteht und in das ein empirisches Relativ abgebildet werden kann.

Die erste Bedingung muß NOTWENDIG erfüllt sein, um Meßmodelle überhaupt empirisch anwenden zu können. Die zweite Bedingung ist eine HINREICHENDE Bedingung, die stets trivial erfüllbar ist, da sich IMMER irgendein, für eine bestimmte Abbildung wünschenswertes, numerisches Relativ finden läßt. Die eigentliche Leistung der Meßtheorie besteht nun darin, daß sie zeigen kann, unter WELCHEN Annahmen (Axiomen), ein empirisches Relativ homomorph (strukturell eindeutig/ relationstreu) in ein numerisches Relativ abgebildet werden kann. Ein solches Axiomensystem definiert eine MESSTRUKTUR, deren empirische Erfüllbarkeit für ein empirisches Relativ hinreichend für den oben geforderten Nachweis ist, daß ein, über die Merkmalsträger in das empirische Relativ eingehendes Merkmal meßbar ist (REPRÄSENTATIONSTHEOREM). Ist die Meßtruktur erfüllt, dann ist zugleich sichergestellt, daß das oder die Merkmale auf einem bestimmten Skalenniveau meßbar sind (EINDEUTIGKEITSTHEOREM). Charakteristisch für einen Skalentyp sind jeweils die zulässigen Transformationen, die die Abbildung des empirischen Relativs im numerischen Relativ nicht zerstören. Von einer "Messung" wird in der Regel erst dann gesprochen, wenn mindestens lineare Skalentransformationen zulässig sind, d.h. wenn eine Intervallskala vorliegt. Gleichzeitig ist damit das sogenannte BEDEUTSAMKEITSPROBLEM gelöst, das sich auf die Frage bezieht, WELCHE numerischen Operationen mit den Daten zu bedeutungsvollen Aussagen führen. Kriterium der Bedeutsamkeit einer numerischen Aussage ist, daß ihre Gültigkeit unter zulässigen Transformationen erhalten bleibt. So wäre beispielsweise die Aussage "die Differenz zwischen einem IQ von 90 und einem IQ von 100 ist gleich der Differenz zwischen einem IQ von 120 und 130" eine formal bedeutsame Aussage, wenn die Testwerte auf einer Intervallskala liegen, sie wären nicht bedeutsam, lägen die Test-

werte auf einer Ordinalskala.

> "Numerische Relationen zwischen den Messungen, welche einer Entsprechung in der Empirie (im empirischen Relativ) ermangeln, sind hingegen inhaltsleer und werden in der Meßtheorie als `sinnlos` bezeichnet" (FISCHER, 1974, S. 118).

Im Zusammenhang mit der Messung LATENTER Verhaltensdispositionen ist ferner die in der Meßtheorie von SUPPES und ZINNES (1963) getroffene Unterscheidung "fundamentaler" und "abgeleiteter" Messungen von Bedeutung. Jede mindestens homomorphe Abbildung eines empirischen Relativs in ein numerisches Relativ konstituiert eine FUNDAMENTALE MESSUNG. Fundamentale Meßvorgänge liegen beispielsweise bei der Bestimmung der Masse und des Volumens eines Körpers, der absoluten Häufigkeit eines beobachteten Verhaltens oder der Gesamtzahl der Beobachtungen vor. Von ABGELEITETER MESSUNG sprechen SUPPES und ZINNES, wenn dem Ergebnis einer oder mehrerer fundamentaler Messungen mit Hilfe einer spezifischen Zuordnungsvorschrift neue Zahlenwerte zugewiesen werden.

"The classical example of a derived measurement is that of density defined as the ratio of mass to volume" (1963, S.17), und ein weiteres, die Bestimmung der relativen Häufigkeit des beobachteten Verhaltens als Quotient aus der absoluten Häufigkeit des Verhaltens und der Gesamtzahl der Beobachtungen. Im Gegensatz zur klassischen MESStheorie (CAMPBELL, 1928), die für abgeleitete Messungen noch strengere Voraussetzungen als für fundamentale Messungen fordert, so z.B. die Möglichkeit der "concatenation operation" (vgl. FISCHER, 1974, S.119), sehen SUPPES und ZINNES (1963) keinen Unterschied zwischen beiden Formen der Quantifizierung. Die Unterscheidung ist dennoch zweckmäßig, wenn man sich ausschließlich auf INDIREKTE QUANTIFIZIERUNGEN zur Messung latenter Dispositionen bezieht und die beiden Begriffe entsprechend umdefiniert:

> "Fundamentale Messung liegt (...) vor, wenn der latenten Variable (...) der numerische Wert der manifesten Variable zugeordnet wird; wenn beispielsweise die mathematische Begabung eines Probanden durch die Anzahl gelöster Aufgaben des entsprechenden Tests ausgedrückt wird. (...) Von abgeleiteter Messung spreche ich, wenn die numerischen Größen der manifesten Variable nur als Basis für die Gewinnung neuer numerischer Werte dienen, die dann als Maßgrößen der latenten Variable bezeichnet werden. (...) Meßmodelle, die zu fundamentalen Meßwerten führen, sollen DATENMODELLE genannt werden. Meßmodelle, die mit abgeleiteten Maßgrößen arbeiten, will ich als PARAMETERMODELLE bezeichnen." (SPADA, 1969a, S.26).

Messen besteht bei den Parametermodellen darin, unbekannte Parameter von Verteilungen aus den Daten einer Stichprobe zu SCHÄTZEN (vgl. FISCHER, 1968, S.79).

Die bereits angesprochene Zirkularität "traditioneller" faktorenanalytischer Modellbildung und Konstruktvalidierung ist wohl letztlich darauf zurückzuführen, daß die Faktorenanalyse, ebenso wie die Klassische Testtheorie (KTT), die Meßbarkeit von Intelligenzmerkmalen auf Intervallskalenniveau -per fiat- voraussetzt, niemals aber wirklich überprüft hat. Die klassische Testtheorie ist ein tautologisches System von Definitionen und Sätzen (LORD und NOVICK, 1968, S.47; FISCHER, 1974, S.114), die nichts über empirisch vorfindbare Ereignisse aussagen. Wegen ihrer geringen Spezifität sind sie auf alle Datenmatrizen anwendbar, die Chance der Falsifikation ist also nicht gegeben. Von daher ist es schwer vorstellbar, wie die KTT und die Faktorenanalyse mit Hilfe von Intelligenztests zu sogenannten Intelligenzdimensionen kommen will, wenn deren Merkmalstruktur auf Intervallskalenniveau nicht meßbar ist.

Die Verwendung von Testwerten als MASSzahlen wird häufig gerechtfertigt mit dem "pragmatischen Anliegen, ein Kriterium aufgrund einer gemessenen Eigenschaft vorherzusagen" (FISCHER, 1970, S. 214). COOMBS et al. (1975) wenden dagegen ein, daß es sich dabei um Prognosen aufgrund "blinder" (S.31) Verfahren handele und HILKE (1980, S.143) gibt schließlich zu bedenken, daß "sich die mit Hilfe von Testscores getroffenen Aussagen auch bei Berücksichtigung der jeweiligen Entscheidungssituation nicht in dem Maße bewährt (haben), daß mit der praktischen Bewährung die Messung mit Tests allgemein gerechtfertigt werden könnte." (Ausführliche Kritiken der KTT und ihrer Verallgemeinerungen finden sich u.a. bei LORD und NOVICK, 1968; SPADA, 1969b; FISCHER, 1974; SCHEIBLECHNER, 1974; LUMSDEN, 1976.)

Das methodische Vorgehen bei der Aufstellung von Axiomen in der Meßtheorie (vgl. ORTH, 1974) erlaubt nun prinzipiell einige Forderungen zu berücksichtigen, die an formal befriedigende Quantifizierungen latenter Dimensionen gestellt werden. Einige der Bedingungen, die als Grundannahmen in das von GEORG RASCH (1960) entwickelte dichotome (spezielle) logistische Parametermodell eingehen, seien im folgenden kurz skizziert. Dabei geht es weniger um die anderenorts ausführlich dargestellte mathematische Herleitung des nach seinem Autor benannten Meßmodells (vgl. RASCH, 1960; FISCHER, 1974), als vielmehr um den Versuch, Implikationen zentraler Modellannahmen im Hinblick auf die Theoriebildung und vor allem die theoretisch angeleitete Aufgabenkonstruktion darzustellen und zu bewerten.

3.1 SPEZIFISCHE OBJEKTIVITÄT

Von wissenschaftstheoretischen Überlegungen zum Begriff der Objektivität von Aussagen ausgehend, kommt RASCH (1966) zu dem Ergebnis, daß "Wissenschaftlichkeit nicht dadurch charakterisiert (ist), daß man überprüfbare Tatsachenfeststellungen macht, sondern daß VERGLEICHE ZWISCHEN VERSCHIEDENEN OBJEKTEN (innerhalb einer Klasse zusammengehöriger Objekte) durchgeführt werden, welche 'ojektiv' sein sollen" (FISCHER, 1974, S.408). Quantitative oder qualitative Vergleiche von Objekten bezüglich einer Eigenschaft sind wissenschaftlich oder SPEZIFISCH OBJEKTIV, wenn sie weder von den "Agentien", mit denen sie in Kontakt gebracht werden, noch von anderen "Objekten" außer den beiden Vergleichsobjekten abhängen. Bezogen auf die Quantifizierung latenter Variablen heißt das: der metrische Vergleich zweier Personen (=Objekte) ist nur dann wissenschaftlich, wenn er unabhängig davon ist, welche Items ("Agentien") einer homogenen Grundmenge vorgelegt wurden und von welchen anderen Personen diese Aufgaben bearbeitet wurden.

Im ersten Absatz des Vorworts zu der Arbeit, in der RASCH (1960) seine grundlegenden meßtheoretischen Überlegungen niederlegt, zitiert er eine Kritik B.F. SKINNERS zur Anwendung KONVENTIONELLER statistischer Verfahren in der psychologischen Forschung:

> "The order to be found in human and animal behavior should be extracted from investigations into individuals (...) psychometric methods are inadequate for such purposes since they deal with groups of individuals" (SKINNER, 1956, S. 221).

Gruppenbezogene Statistiken, wie sie in der traditionellen psychometrischen Praxis verwendet werden (Faktorenanalyse, Varianzanalyse etc.), erfordern die Zusammenfassung von Individuen in diskrete Klassen oder Populationen und treffen zudem Annahmen über die Natur der Variabilität innerhalb dieser Kategorien. RASCH teilt die Zweifel über die Angemessenheit dieser Vorgehens-

weisen und Annahmen: eine Alternative sah er allein darin, PERSONENbezogene Meßmethoden zu entwickeln:

> "Individual-centered statistical techniques require models in which each individual is characterised separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments, tests, or items or other stimuli within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class -measuring the same thing- independent of which particular individuals within the class considered were instrumental for the comparison" (RASCH, 1960, VII).

Mit diesem Exkurs greift RASCH eine Thematik auf, die schon von L.L. THURSTONE drei Jahrzehnte zuvor entwickelt worden war:

> "A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extend that its measurement function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, than to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instruments is intended its function must be independent of the objects of measurement" (THURSTONE, 1928, S.547).

Diese Analogie zu physikalischen Messungen ist charakteristisch für die Argumentationsweise vieler Latent-Trait Theoretiker. So bemühen beispielsweise auch WRIGHT (1968, 1977) und WRIGHT & STONE (1979) den "yardstick" als Metapher für das Testitem. Andere verweisen auf Ähnlichkeiten zwischen der Messung laten-

ter, mentaler Dispositionen und der Temperaturmessung. Dabei liegt stets die Annahme zugrunde, das psychologische Messungen zwar schwieriger zu realisieren seien als Messungen im Bereich der Physik oder Chemie, nichtsdestoweniger sollten sie jedoch den gleichen generellen Prinzipien folgen. RASCH selbst vergleicht die Entwicklung seiner Modelle mit MAXWELL's Analysen der NEWTON'schen "Bewegungsgesetze" in Begriffen der Konzepte "Kraft" und "Masse". Das "2. Gesetz " setzt "Kraft", "Masse" und "Beschleunigung" miteinander in Beziehung und zwar in einer Situation, in der die "Beschleunigung" und ihre Messung zwar vollständig diskutiert, die Konzepte "Masse" und "Kraft" jedoch nicht definiert sind. RASCH (1960, S.110-114) betrachtet es als notwendig beide Konzepte in aufeinander bezogenen Begriffen zu definieren und zeigt, wie Transformationen des Modells ("Gesetzes") und eine entsprechende Datensammlung sowohl (komparative) Messungen von Massen als auch die (komparative) Messung von Kräften gestatten.

Die enge Analogie zum eindimensionalen logistischen Meßmodell, das Personenfähigkeiten, Aufgabenschwierigkeiten und Lösungswahrscheinlichkeiten miteinander in Beziehung setzt, ist offensichtlich: die Konzepte "Personenfähigkeit" und "Aufgabenschwierigkeit" erfordern aufeinander bezogene Definitionen, denn die Disposition bestimmte "Probleme" lösen zu können ist nur aus entsprechenden Aufgabenbeantwortungen erschließbar und die Schwierigkeit eines Items zeigt sich erst, wenn es von Personen bearbeitet worden ist. Spezifisch-objektive, d.h. wechselseitig stichprobenunabhängige, komparative Messungen von Personenfähigkeiten und Aufgabenschwierigkeiten sind möglich, wenn Antwortdaten vorliegen, die separierte Schätzungen der die Personenfähigkeiten und Itemschwierigkeiten charakterisierenden Modellparameter erlauben. Sind Personen- und Itemparameter in diesem Sinne verteilungsunabhängig schätzbar, dann ist sichergestellt,

> "a) daß die auf eine (...) Person bezogenen (...) Aussagen weder von der willkürlichen Zuordnung der

> Person zu einer bestimmten Referenzpopulation abhängen, noch von der mehr oder weniger beliebigen Auswahl von Testitems (...);
> b) daß Veränderungen von psychologischen Leistungsdimensionen unabhängig vom Ausgangsniveau der Person und unabhängig von den verwendeten Testaufgaben beurteilt werden können;
> c) daß (...) die sowohl theoretisch als auch praktisch problematische Repräsentativität der Stichprobe völlig an Bedeutung verliert" (FISCHER, 1974, S.407).

Mit der ZENTRALEN Forderung nach spezifisch objektiven Quantifizierungen latenter Verhaltensdimensionen sind eine Reihe wichtiger Implikationen verbunden.

Die erste Forderung betrifft die Homogenität der Aufgaben bzw. die Eindimensionalität der angesprochenen latenten Fähigkeit. Parametermodelle, wie z.B. das RM, definieren "Homogenität" - anders als etwa die KTT- verteilungsunabhängig als eine Eigenschaft des abzubildenden Maßsystems. Bei Beschränkung auf zweikategoriale Items wird der Homogenitätsbegriff hier durch das Prinzip der lokalen stochastischen Unabhängigkeit der Reaktionen formalisiert. Es besagt, daß sich die Wahrscheinlichkeit, mit der eine Person einen bestimmten Antwortvektor liefert, als Produkt der Einzelwahrscheinlichkeiten ergeben muß, mit denen die Person die einzelnen Aufgaben löst bzw. nicht löst, oder anders ausgedrückt: die Beantwortung einer Aufgabe beeinflußt in keiner Weise die Antworten auf nachfolgende Aufgaben. Wäre dies nicht der Fall, würde jede Veränderung der Aufgabenreihenfolge den Test als Meßinstrument und damit die Lokalisation der Personen auf der latenten Dimension verändern. Ist die Modellannahme hingegen erfüllt, dann ist sichergestellt, daß Personen und Aufgaben durch je einen Parameter charakterisiert werden können, d.h. die Aufgaben sind homogen, da sie Indikatoren einer einzigen latenten

Fähigkeit sind und ebenso sind die Personen gegenüber den Aufgaben homogen, da für sie die Itemschwierigkeit durch EINEN Parameter ausdrückbar ist.

RASCH (1960) geht weiter von der allgemeinen Annahme stetiger latenter Dimensionen der Personen (ξ bzw. $\ln(\xi) = \Theta$) und der Items (σ bzw. $\ln(\sigma) = b$) aus, d.h. jede Person, jedes Item und jede Person x Item- Kombination (=Situation) sollen im Hinblick auf den angestrebten Vergleich durch jeweils genau einen Parameter vollständig beschreibbar sein. Dabei wird gefordert, daß alle Parameter auf EINER Dimension lokalisiert sind und daß sich der -als "Reaktionstendenz" oder "Reaktionsstärke" interpretierbare- "Situations-"parameter als Produkt des Personen- und Itemparameters ergeben soll. Während der latente Situationsparameter deterministisch von der Personenfähigkeit und der Aufgabenschwierigkeit abhängt, sollen Zufallseinflüsse bei der Beantwortung durch Angabe einer stetigen (hier: logistischen) Wahrscheinlichkeitsverteilung möglicher Antworten bei gegebenen Parametern berücksichtigt werden. BIRNBAUM (1968) zeigt, daß aus den aufgeführten Annahmen und bei Beschränkung auf zweikategoriale Items, mathematisch notwendig die Itemcharakteristik (ICC) f(vg) des RM folgt:

$$(3) \qquad f_{vg} = p(x_{vg} = 1 | v,g)$$

$$= [1 + \exp(b_g - \Theta_v)]^{-1}$$

Durch die Annahme lokal stochastisch unabhängiger Aufgabenbeantwortungen und Gleichung (3) ist das zweifaktorielle bzw. zweikategoriale Testmodell von RASCH (1960) festgelegt. Die Gleichung (3) ist eine Abbildungsfunktion, die (nur) bei Geltung der aufgeführten Bedingungen einem abgeleiteten Maßsystem, bestehend aus den Wahrscheinlichkeiten p neue Werte Θ für alle Personen und neue Werte b für alle Aufgaben zuordnet und damit ein numerisches Relativ konstituiert. Eine itemcharakteristische Kurve beschreibt also, wie wahrscheinlich es ist, daß eine Person v mit der latenten Fähigkeit Θ eine Aufgabe g der Schwierigkeit b richtig löst.

Item- und Personenparameterschätzungen liegen, bezogen auf Gleichung (3) auf einer Differenzenskala, d.h. Meßwertdifferenzen sind eindeutig bis auf eine durch Normierung festzulegende Konstante c.

Die geforderte Eigenschaft der spezifischen Objektivität ist eng verknüpft mit der Möglichkeit, Personen- und Itemparameter bei der Schätzung zu trennen. Voraussetzung für die Schätzung der Itemparameter bei Unkenntnis der Personenparameter (und vice versa) ist die Existenz eines Satzes (minimal) erschöpfender Statistiken für Item- und Personenparameter (vgl. ANDERSEN, 1977).. Ein Satz solcher Statistiken für diese Parameter liegt in Form der Itemrandsummen x(.g) und der Personenrandsummen (Testscores) x(v.) vor. Das RASCH Modell ist das einzige probabilistische Testmodell, in dem die Randsummen (x(.g), x(v.)) alle in den Daten über Personen- und Itemparameter enthaltene Information repräsentieren (vgl. dazu BIRNBAUM, 1968; FISCHER, 1974). Stichprobenunabhängige Schätzungen der Itemparameter sind nun prinzipiell möglich, da die bedingte Wahrscheinlichkeit (conditional likelihood) aller Itemrandsummen x(.g) eine Funktion ist, die NUR von den Itemparametern b(g) abhängt (für eine detaillierte Darstellung und Beweisführung vgl. FISCHER, 1974, S.221ff.).

$$(4) \qquad p\,[(x_{.g} \mid x_{v.})] = f\,[(x_{v.}), (x_{.g}), (b_g)]$$

Nach der Maximum-Likelihood-Methode werden obige bedingte Wahrscheinlichkeit durch die Wahl geeigneter Itemparameter b(g) maximiert. Durch das "conditional maximum likelihood" (CML-) Verfahren gelangt man zu Schätzgleichungen, die iterativ lösbar sind (vgl. Kp. 5.1 und FISCHER, 1974, S.230ff.).

Von entscheidender Bedeutung ist, daß durch das Prinzip der "bedingten Inferenz" (ANDERSEN, 1973) statistische Information, die "über die Items in den Daten enthalten ist, (...) ohne Berücksichtigung der Fähigkeit der getesteten Personen ausgewer-

tet werden kann" (FISCHER, 1974, S.233). Durch diese vorteilhafte Eigenschaft der Parameterschätzung sind allgemeingültige Aussagen auch für den diskutierten Fall "partiell konsistenter Beobachtungen" möglich.

> "Eine Folge der Struktur dieser Modelle ist die 'Stichprobenunabhängigkeit', d.h. nur bei Verwendung dieser Modelle erhält man Schätzungen von Situations- (Reiz- oder Reaktions-) Parametern, die unabhängig von der verwendeten Vpn-Stichprobe sind" (SCHEIBLECHNER, 1972, S.476f.).

Ungeachtet der weithin unbestrittenen meßtheoretischen Vorzüge dieses sparsamen Meßmodells, sollte nicht übersehen werden, daß die Diskussion über seine Anwendbarkeit in der psychologischen und pädagogischen Testdiagnostik, gerade im Hinblick auf die strengen Annahmen des Modells, kontrovers geführt wird. So sind einmal jene Autoren zu nennen, deren Kritik sich generell dagegen richtet, Latent Trait Modelle in der Leistungsmessung einzusetzen. So betont beispielsweise GOLDSTEIN (1979) den fundamentalen Unterschied zwischen "well-ordered-traits", wie "aptitude" und "intelligence" auf der einen Seite und den komplexen Verhaltensmustern des "educational achievement" auf der anderen Seite. Seiner Auffassung nach MUSS jedes eindimensionale Messmodell im Bereich der Leistungsmessung scheitern.

Daß dieser Vorbehalt so allgemein sicherlich nicht gerechtfertigt ist, zeigen erfolgreiche Anwendungen des RASCH-Modells in der Schulleistungsmessung und -diagnostik z.B. bei FRICKE (1972), SPADA, HÄUSSLER und HEYNER (1973), HÄUSSLER (1981) und WRIGHT und BELL (1984).

Unter den Befürwortern latenter Testmodelle in der pädagogischen und psychologischen Testdiagnostik konzentriert sich die Diskussion auf die Frage, ob das "einparametrige" logistische Modell von RASCH oder komplexere, mehrparametrige Testmodelle, wie das in den USA besonders populäre 3-Parametermodell von BIRNBAUM

(vgl. LORD und NOVICK, 1968), geeigneter sind, das Reaktionsgeschehen bei der Begegnung eines Individuums mit einer bestimmten Aufgabe zu modellieren. Wenn diese Frage zugunsten mehrparametriger eindimensionaler oder gar mehrdimensionaler Modelle entschieden wird, so geschieht dies zumeist auf der Grundlage der empirischen Modellbewährung. Damit ist das Ausmaß angesprochen, in dem die beobachteten Testdaten den Annahmen des zugrundegelegten Meßmodells entsprechen. GUSTAFSSON (1981) bemerkt jedoch, daß der Modellfit kein absolutes Entscheidungskriterium für die Wahl eines Testmodells sein kann, sondern stets im Lichte des jeweiligen Anwendungsfalles bewertet werden sollte. EINE typische Anwendung besteht darin, existierende, mit traditionellen Testverfahren erhobene Daten mit Hilfe eines latenten Testmodells zu analysieren. In diesen Testmaterialien befinden sich in der Regel zahlreiche Aufgaben, die keine spezifisch objektiven Messungen erlauben, da sie beispielsweise Assoziationen provozieren oder nicht bei allen Personen einer Population dasselbe Merkmal ansprechen. Anhänger des RASCH-Modells befürworten die Selektion modellinkonformer Items (vgl. SCHEIBLECHNER, 1974; FISCHER, 1974) oder vertrauen auf die Robustheit des Modells, d.h. auf die für praktische Meßprobleme mitunter vernachlässigbare Beeinträchtigung des Meßergebnisses durch Verletzungen der Modellannahmen (GUSTAFSSON, 1980; WAINER und WRIGHT, 1980). Konsequenter scheint unter solchen Umständen aber, eine merkmalsbezogene Testkritik anzustreben (DIVGI, 1981), oder aber ein Meßmodell anzupassen, das bestimmte Defekte, etwa durch die Annahme eines Rateparameters oder die Möglichkeit überschneidender itemcharakteristischer Kurven, modelliert (vgl. z.B. WEISS und DAVISON, 1981; BEJAR, 1983).

Bei der Planung NEUER Tests oder Itembanken kann das Konstruktionsziel aber nur darin bestehen, meßtheoretisch brauchbare, d.h. spezifisch objektive Instrumente zu entwickeln. In diesem Kontext fungieren strenge Modellkontrollen zum RM als Kriterien für die Eindimensionalität des gemessenen Merkmals und die Homo-

genität der Aufgaben.

> "It is obvious, that when the RASCH modell is used as a criterion, high standards of fit must be set and it is necessary that several tests which each guard against different deviations from the modell are applied" (GUSTAFFSON, 1980, S.225).

CHOPPIN (1984) unterscheidet ebenfalls diese beiden Anwendungszusammenhänge von Latent-Trait-Modellen und plädiert für die Anwendung des strengeren und erklärungskräftigeren RM in der Itemkonstruktion:

> "Those whose main interest is in understanding existing data sets, and therefore modeling of observed ICCs, see little benefit or relevance in specific objectivity. Those who wish to construct instruments to measure individuals optimally tend to prefer the approach which offers the stronger and more useful system. ICCs which reflects the behavior of inefficient or ineffective items have little interest for them. (...) The RASCH modell supports a range of applications which goes well bejond what a latent trait model that is not specifically objective can manage" (S.28).

Einen weiteren Vorteil des Modells sieht CHOPPIN (1984) im Zusammenhang mit ökonomischen Überlegungen zur Konstruktion umfangreicher Itembanken:

> "An item bank requires calibration, and although in theory there are alternative approaches, in practice the RASCH modell has proved by far the most cost effective and is the most widely used" (1984, S.22).

Die rechentechnisch aufwendige Kalibrierung der Parameter komplexer Modelle erfordert oft mehrere tausend Antwortvektoren, während die Schwierigkeitsparameter des RM mit einfachen Algorithmen meist schon aus 200 oder weniger Responsevektoren zufrie-

denstellend geschätzt werden können (vgl. WRIGHT und STONE, 1979; LORD, 1983; siehe auch Kp. 5.1).

Wie schon dem Zitat von GUSTAFSSON (1980) zu entnehmen ist, kommt den Modellkontrollen gerade für den hier interessierenden Bereich der Test- und Aufgabenkonstruktion zentrale Bedeutung zu. Die Vorteile, die sich aus der Anwendung von Item-Response Theorien ergeben können, greifen nicht, wenn der Fit zwischen dem Modell und den Testdaten unzureichend ist. Im Vergleich etwa zu Fragen der Modellbildung und Parameterschätzung, wurde den methodischen Problemen der Modellbewährung in konkreten Anwendungssituationen vergleichsweise wenig Aufmerksamkeit zuteil (vgl. HAMBLETON und MURRAY, 1983). Selbst Anwender des RM, für das strengere und vielfältigere Gültigkeitskontrollen als für andere gängige latente Testmodelle existieren, konzidieren diesen Schwachpunkt. Zwar wurden gerade für dieses Modell in den letzten Jahren Teststatistiken entwickelt (vgl. GUSTAFSSON, 1980), ihre Aussagekraft ist aber eingeschränkt, da sie stark von dem Umfang der jeweiligen Aufgaben- und Personenstichprobe abhängig sind.

> "When sample sizes are large, nearly all departures between a modell and a data set (even those where the practical significance of differences is minimal) will lead to rejection of the null hypothesis of modell data fit. With small sample sizes even big differences may NOT be detected via statistical methods because of the low level of statistical power" (HAMBLETON und MURRAY, 1983, S.72).

Mangelnde Übereinstimmung besteht insbesondere zwischen denen, die Fit bzw. Misfit als dichotome Kategorisierung auffassen und Testdaten auf dieser Grundlage beurteilen und verarbeiten (z.B. GUSTAFSSON, 1980) und jenen, die das Ausmaß des Misfit als kontinuierliche Variable interpretieren, die im Kontext der gesamten Situation auch weiterführende Informationen über mögliche Ursachen mangelnder Modellverträglichkeit für einzelne Items UND

Probanden liefern kann (z.B. WRIGHT und STONE, 1979). Das Unterkapitel 5.3 diskutiert Zugänge zu multiplen Gültigkeitskontrollen im Zusammenhang mit der Konstruktion und Evaluation eines umfangreichen Aufgabenpools.

Das RM stellt in meßtheoretisch befriedigender und prüfbarer Weise die geforderte, potentiell falsifizierbare Beziehung zwischen "person performance" und "item characteristics" her. Damit kann die von EMBRETSON (1983) unter Punkt 1 genannte Teilforderung (vgl. S.40) einer formal adäquaten Konstruktrepräsentation als erfüllt gelten. Das Modell hat jedoch in dieser Ausgangsformulierung, wie jedes ItemANALYSEmodell, keinen substanztheoretischen Bezug. Um beobachtete Testdaten, unter Beibehaltung aller meßtheoretischen Vorzüge des RM, im Sinne inhaltlich- theoretischer Vorstellungen über das gemessene latente Merkmal interpretieren zu können, sind präexperimentelle Annahmen über kognitive/ affektive Lösungskomponenten mit Hilfe oder in Verallgemeinerungen bzw. Spezifikationen des Modells abzubilden.

4. LOGISTISCHE AUFGABENBEARBEITUNGSMODELLE

Geht man einmal von den im Zusammenhang mit der Komponentenanalyse (vgl. Kp. 2.3.1) vorgestellten Methoden, relevante Informationsverarbeitungskomponenten eines bestimmten Aufgabentyps zu identifizieren aus, so stehen mindestens zwei Zugänge für die Modellierung mentaler Lösungsoperationen im Rahmen der RASCH-Meßtheorie offen:

1. Die Subtask-Methode unterteilt eine Aufgabe, nach informations- oder kognitionstheoretischen Gesichtspunkten, in mehrere unabhängige oder sequentiell abhängige Komponenten. Experimentelle Designs, wie sie oben beschrieben wurden (vgl. Kp. 2.3.1), erlauben für bestimmte Aufgabentypen, Bearbeitungszeiten der angenommenen kognitiven Komponenten oder Daten über deren richtige bzw. falsche Durchführung zu erheben. Ist die Beteiligung einer Operation am Lösungsprozeß identifizierbar, so läßt sich mit Hilfe des RM die Wahrscheinlichkeit, daß eine Person v eine Aufgabe g löst, auf verschiedene Weise modellieren (Kp. 4.1).

2. Üblicherweise liegen Antwortdaten jedoch auf Itemebene vor. Aber auch in diesem Fall können Annahmen über die zur Aufgabenlösung erforderlichen Operationen im Rahmen der RASCH-Meßtheorie formalisiert und empirisch geprüft werden. Die "Methode der Komplexitätsfaktoren" setzt voraus, daß entweder aus einer Theorie des betreffenden Merkmalsbereichs oder mittels struktureller Analysen von Aufgabeninhalten Hypothesen darüber abgeleitet werden können, welche Lösungskomponenten die Schwierigkeit einer Aufgabe bedingen (Kp. 4.2).

4.1 DIE LOGISTISCHEN KOMPONENTENMODELLE VON SPADA UND EMBRETSON

Die in diesem Kapitel vorzustellenden Komponentenmodelle setzen die unter Punkt 1 angesprochene Möglichkeit voraus, Daten über

die (richtige oder falsche) Beantwortung von Aufgaben zu erheben. Da dies nur in seltenen Fällen und unter - möglicherweise (!) verfälschenden - experimentellen Bedingungen möglich sein wird, sollen diese Modelle nur kurz und hauptsächlich unter formalen Aspekten betrachtet werden.

SPADA (1976) geht von der Modellstruktur des probabilistischen Automatenmodells von SUPPES und MORNINGSTAR (1972) mit ausschließlich inzidentellen (individuellen) Parametern aus:

(5) $p\{\text{Person v löst Item g richtig}\}$

$$= \prod_{j=1}^{m} p\{\text{Person v führt Op. j richtig durch}\}^{f_{gj}}$$

Durch Einsetzen der logistischen Funktion des RM zur Beschreibung von Operationslösungswahrscheinlichkeiten durch individuelle (inzidentelle) und allgemeine (strukturelle) Parameter, erhält SPADA (1976, S.116) die Grundgleichung eines Modells, das er, seiner Herleitung entsprechend, als "logistisches Automatenmodell" (LAM) bezeichnet:

$$(6)\quad p\{x_{vg} = 1 \mid \theta_v, \alpha_j\} = \prod_{j=1}^{m} [1/\{1 + \exp(\alpha_j - \theta_v)\}]^{f_{gj}}$$

Das RM beschreibt hier, analog zur oben diskutierten Interpretation des RM für Aufgabenlösungswahrscheinlichkeiten, die bedingte Wahrscheinlichkeit mit der eine Person v mit der Fähigkeitsausprägung θ_v eine Operation j der Schwierigkeit α_j ausführt. Das LAM verknüpft diese Ausführungswahrscheinlichkeiten multiplikativ zu einer theoretisch erwarteten Lösungswahrscheinlichkeit p(x(vg)) für die Gesamtaufgabe. SPADA (1976, S.116) sieht den besonderen Vorteil des LAM gegenüber dem Modell von SUPPES darin, "daß das Problem individueller Unterschiede weder negiert noch zum Anlaß einer nicht handhabbaren Parameterinflation gemacht wird". Jede Person v ist in ihrer Fähigkeit Operationen des betreffenden Typs ausführen zu können, durch genau einen Parame-

ter θ_v charakterisiert. Das EINdimensionale LAM nimmt ferner an, daß die Schwierigkeit einer Operation j unabhängig davon ist, in welcher Aufgabe oder an welche Stelle eines Lösungsalgorithmus die Operation durchzuführen ist. Die wichtige Annahme KONTEXTUNABHÄNGIGER Operationsschwierigkeiten ist gezielt empirisch prüfbar, indem die aus "Operationsdaten" geschätzten Schwierigkeitsparameter für eine bestimmte Operation auf ihre modellseitig angenommene statistische Äquivalenz untersucht werden. Kontext- oder materialunabhängige Operationsschwierigkeitsschätzungen stellen eine wichtige Voraussetzung für die Ableitung einer Itemkonstruktionstechnologie aus der betreffenden Konstruktrepräsentation dar. Variieren die Operationsschwierigkeiten in unbekannter Weise mit nicht weiter spezifizierten Itemcharakteristika (z.B. Materialfaktoren, Position in der Itemreihe o.ä.), so ist eine rationale, meßzielorientierte Synthese konstruktvalider Aufgaben nicht möglich.

An dieser Stelle sei noch erwähnt, daß in das LAM die interessante Annahme eingeht, die angenommenen Operationen seien in dem Sinne elementar, daß sie nicht auf weitere Teiloperationen zurückgeführt werden können. Angenommen für eine Operation α gelte

$$p\,(x_{\alpha,v} = 1) = [1/(1 + \exp(\alpha_1 - \theta_v))][1/(1+\exp(\alpha_2 - \theta_v))], \text{ für alle } v \tag{7}$$

dann ließe sich für die aus den Teiloperationen α_1 und α_2 zusammengesetzte Operation α kein Schwierigkeitsparameter schätzen, der für alle Personen v gültig wäre. Statt dessen wären Daten über die Operationen α_1 und α_2 zu erheben. SPADA (1976, S.118) macht darauf aufmerksam, daß mangelnde Modellgültigkeit darauf zurückzuführen sein kann, daß durch inadäquate Hypothesen über den Lösungsprozeß, Daten zu nicht elementaren Operationen erhoben worden sind. Ergänzend sei angemerkt, daß auch ein unangemessenes oder nicht hinreichend kontrolliertes experimentelles

Datenerhebungsdesign dazu führen kann, daß die Annahmen über die Komponenten des Lösungsprozesses nicht bestätigt werden, obwohl sie möglicherweise zutreffen.
Liegen binäre Beobachtungsdaten auf Operationsniveau vor, bereitet die Anwendung des LAM (Parameterschätzungen und Gültigkeitskontrollen) keine BESONDEREN Probleme, da es sich hierbei lediglich um eine Realisation des RM für Operationslösungswahrscheinlichkeiten handelt. Hier sei auf eine Untersuchung von SPADA (1976) hingewiesen, in der er das LAM in modifizierter Form (Modellierung von Lernprozessen durch den Verzicht auf die Annahme zeitlich konstanter Operationsparameter) erfolgreich bei der Analyse von Drehrichtungsaufgaben anwendet (S.254 ff.).

WHITELY (1980; jetzt EMBRETSON) schlägt ein "Multicomponent Latent Trait Modell" (MLTM) vor, das als mehrdimensionale Verallgemeinerung des LAM aufgefaßt werden kann. Analog zum eindimensionalen LAM dienen Beobachtungsdaten über die erfolgreiche bzw. erfolglose Bearbeitung von Subtasks, die -idealiter- theoretisch postulierte, elementare Lösungskomponenten (= Operationen) repräsentieren sollen, zur Schätzung von Komponentenschwierigkeiten für einzelne Komponenten/ Operationen und der Lösungswahrscheinlichkeit für das gesamte Item.

Die Komponenten können dabei, je nach Modellformulierung, als voneinander unabhängig oder sequentiell abhängig angenommen werden. Im ersten Fall hat die Bearbeitung einer Subtask keinerlei Einfluß auf nachfolgende kognitive Operationen, während im zweiten Fall eine Operationsdurchführung notwendige Informationen für die erfolgreiche Ausführung nachfolgender Operationen bereitstellt.

Gleichung (8; vgl. EMBRETSON, 1984) stellt ein mathematisches Modell für UNABHÄNGIGE Komponenten dar, indem die Lösungswahrscheinlichkeit für die Gesamtaufgabe gT als Produkt der Komponentenlösungswahrscheinlichkeiten aufgefaßt wird:

$$(8) \qquad p(x_{vgT} = 1) = a \prod_{K} p(x_{vgK} = 1) + g[1 - \prod_{K} p(x_{vgK} = 1)]$$

wobei:

$p(x_{vgT} = 1)$:= Lösungswahrscheinlichkeit für die Gesamtaufgabe

$p(x_{vgK} = 1)$:= Wahrscheinlichkeit, mit der Person v die Operation/ Komponente K in Item g richtig ausführt

a := Wahrscheinlichkeit für die richtige Beantwortung des Gesamtitems, wenn die Operationen alle korrekt durchgeführt werden

g := Wahrscheinlichkeit für die richtige Beantwortung des Gesamtitems, wenn mindestens eine Subtask falsch beantwortet wird

In Erweiterung des ursprünglich vorgeschlagenen MLTM (s. WHITELY, 1980), beinhaltet diese Formulierung zwei Konstanten, "a" und "g", wobei die letztere eine gemittelte Ratewahrscheinlichkeit bei der Gesamtitemlösung und "a" den durchschnittlichen Einfluß einer Metakomponente beschreibt, die man als "exekutive Kontrollfunktion" charakterisieren könnte (zur Schätzung von "a" und "g" vgl. EMBRETSON, 1984, S.181). Ihre Funktion und Beziehung zur Gesamt- und Komponentenbeantwortung wird deutlicher, wenn man die, durch das mathematische Modell beschriebene, Struktur des "probability space" betrachtet. Bezeichnet man mit K die Anzahl hypostasierter Komponenten/ Subtasks, so sind 2**K+1 verschiedene Antwortvektoren möglich. Tabelle 1 zeigt die acht möglichen Antwortvektoren für ein Zwei-Komponten-Item und ihre Auftretenswahrscheinlichkeiten. Die Parameter "a" und "g" setzten die Komponentenbeantwortungen mit der Gesamtitembeantwortung in Beziehung, während die übrigen Terme die Wahrscheinlichkeit der Komponenten-Antwortpattern beschreiben.

C(1)	C(2)	T	Notation
1	1	1	a P(x1) P(x2)
1	1	0	(1-a) P(x1) P(x2)
0	1	1	g Q(x1) P(x2)
0	1	0	(1-g) Q(x1) P(x2)
1	0	1	g P(x1) Q(x2)
1	0	0	(1-g) P(x1) Q(x2)
0	0	1	g Q(x1) Q(x2)
0	0	0	(1-g) Q(x1) Q(x2)

Tabelle 1:
Bedingte Wahrscheinlichkeiten möglicher Antwortvektoren
bei 2-Komponenten-Items

Es kann gezeigt werden (vgl. EMBRETSON, 1984), daß die Summe der Wahrscheinlichkeiten für die acht Antwortvektoren 1.0 beträgt, wobei EMBRETSON (1984) die bedingte Aufgabenlösungswahrscheinlichkeit (x(T)=1) durch Einsetzen des RM wie folgt spezifiziert:

$$(9) \qquad p\,[x_{vgT} = 1 \mid (\theta_v), (b_g)]$$

$$= (a-g) \prod_K \frac{\exp(\theta_{vK} - b_{gK})}{1 + \exp(\theta_{vK} - b_{gK})} + g$$

wobei $(\theta(v))$:= Vektor der K Komponentenfähigkeiten der Person v
$(b(g))$:= Vektor der K Komponentenschwierigkeiten für Item g

Der wesentliche Unterschied zum LAM besteht nun darin, daß individuelle Fähigkeitsunterschiede bezüglich jeder der postulierten Komponenten/ Operationen parametrisiert werden. Das MLTM erfüllt somit strukturell neben den ersten drei auch die vierte Forderung EMBRETSON's an eine Methodologie der internen Konstruktvalidierung (s. S.40). Motivation und Vorteile der getrennten Parame-

trisierung von Komponentenfähigkeiten faßt EMBRETSON (1982, S.22f.) in vier Punkten zusammen:

> "FIRST the various processing components from which information is required for item solution are theoretically distinct. Experimental cognitive research has supported the independence of components within a task by additive factor (STERNBERG, 1969) or subtractive factor (PACHELLA, 1974) modeling methods.
> SECOND, if the components are sufficiently elementary, they should generalize across tasks and possibly account for differential patterns of correlations in performance on separate types of items (CARROLL, 1974).
> THIRD, individual differences on different components correlate only moderately, and show differential validity in predicting performance on other tasks (STERNBERG, 1977; WHITELY, 1980).
> FOURTH, component difficulties are sometimes not highly correlated in item sets, so that it is possible to select items of the same type that measure DIFFERENT component abilities."

WHITELY (1980) demonstriert eine Validitätsdekomposition am Beispiel differentieller korrelativer Beziehungen der aus Wortanalogie-Subtasks (s. Bsp. auf S.27) geschätzten MLTM-Komponentenfähigkeitsparameter zu verschiedenen Schulleistungstests des American College Testing Program Examination. Auf die methodischen Probleme, N*K-Personenparameter hinreichend reliabel zu schätzen oder auch die Gleichheit einer postulierten Komponente in verschiedenen Aufgabentypen empirisch nachzuweisen, soll nicht näher eingegangen werden. Hier sei auf die ausführlichen Darstellungen EMBRETSON's (WHITELY's, 1980, 1983) hingewiesen.

Gerade im Hinblick auf die technologische Brauchbarkeit des Modells im Rahmen meßzielorientierter Itemkonstruktion ist jedoch ein anderes Problem von grundsätzlicher Bedeutung. Die Verallgemeinerung des LAM zu einem mehrdimensionalen Modell, das KOMPONENTENSPEZIFISCHE Personenfähigkeiten postuliert, setzt zwangsläufig voraus, daß die Annahme kontextunabhängiger Operations- bzw. Komponentenschwierigkeiten zugunsten itemspezifischer Komponentenschwierigkeiten (b(gk)) aufgegeben werden muß. Während das eindimensionale LAM kontextunabhängige Operationsparameter annimmt und diese Annahme empirischen Kontrollen unmittelbar zugänglich macht, ist die Abhängigkeit der Komponentenschwierigkeiten von modellseitig NICHT berücksichtigten oder kontrollierten Aufgabenspezifitäten für das MLTM konstitutiv, da die Komponentenfähigkeiten Θ(vk) nur dann schätzbar sind, wenn die Operationsschwierigkeiten b(gk) Varianz aufweisen.

Abbildung 3 gibt die Schwierigkeitsschätzungen der beiden Komponenten "Rule Construction" und "Response Evaluation" für 45 von EMBRETSON (1982) untersuchte verbale Klassifikationsaufgaben (s. S.28) wieder.

Die Korrelation zwischen den Komponentenschwierigkeiten ist niedrig (.16), so daß im Diagramm Aufgaben mit substantiell unterschiedlichen Komponentenschwierigkeiten erkennbar sind. Aufgrund der geschätzten Schwierigkeitsparameter ist es möglich, Items zu einem Test zusammenzustellen, der die Messung einer spezifischen Komponentenausstattung gestattet. So dürfte beispielsweise die Auswahl von Items, in denen "Response Evaluation(s)" für nahezu alle Pbn hohe Lösungswahrscheinlichkeiten aufweisen, "Rule Constructions" aber "echte" kognitive Anforderungen darstellen (z.B. Items 6, 11, 32, 34 und 36) zu einem Test führen, der primär die Fähigkeit von Personen zur "Rule Construction" mißt. Es scheint somit durchaus möglich, aus einem kalibrierten Itemset, der Aufgaben gleichen Typs enthält, Tests zu entwickeln, die in unterschiedlichem und bestimmbaren Ausmaß

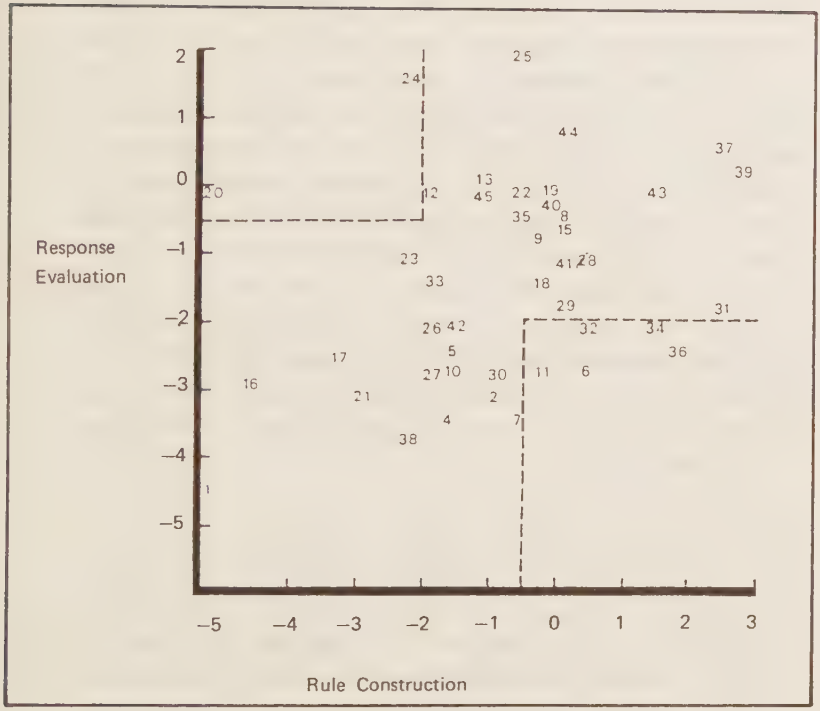


Abbildung 3 :
Komponentenschwierigkeitsschätzungen für
45 verbale Klassifikationsaufgaben
(aus EMBRETSON, 1982)

zur Messung verschiedener Komponentenfähigkeiten geeignet sind. Eine Voraussetzung ist allerdings wieder, daß ausreichend viele Aufgaben und deren Komponentenschwierigkeitsschätzungen zur Verfügung stehen. Dem Itemkonstrukteur stellt sich also wieder das Problem, wie er zu Aufgaben gelangen kann, die den Anforderungen

des Modells genügen. Da das MLTM selbst NICHT spezifiziert, WIE eine Subtask beschaffen sein muß, damit die Wahrscheinlichkeit ihrer Lösung durch das Modell (RM auf Komponentenebene) dargestellt werden kann, und da es ebenfalls nicht beschreibt, WARUM eine bestimmte Komponente für viele Personen in Item x "leicht" und Item y vergleichsweise "schwer" durchzuführen/ lösbar ist, muß das Multicomponent Latent Trait Model für die hier anvisierten Zwecke wohl als weitgehend unbrauchbar betrachtet werden. Auch aus inhaltlich theoretischer Sicht kann die Beschreibung des Lösungsprozesses bei verbalen Klassifikationsaufgaben durch nur zwei recht globale Mechanismen nicht zufriedenstellen. Selbst WHITELY (1981) räumt ein, daß die von ihr postulierten Komponenten Kombinationen mehrerer möglicherweise grundlegenderer Aufgabenlösungsprozesse darstellen; nicht primär aus einer psychologischen Theorie abgeleiteten Hypothesen, sondern die mit Blick auf das experimentelle Vorgehen gegebene Möglichkeit, Subtasks zu bilden, scheinen zu den Definitionen der Komponenten "Rule Construction" und "Response Evalusation" zu führen.

Festzuhalten bleibt, daß die vierte Forderung EMBRETSON's (1983, S. 181), "(...) the measurement of persons on the constructs (...)"; im Rahmen der RASCH-Meßtheorie zu formalen Konstruktrepräsentationen führt, die nicht im intendierten Sinne für rationale Aufgabenkonstruktionen technizierbar sind. Vorausgesetzt, daß Aufgaben eines bestimmten Typs theoriegeleitet in Subtasks unterteilt werden können, deren Bearbeitung singuläre elementare Komponenten bzw. Operationen einer eindimensionalen Lösungsfähigkeit erfordern, kann das eindimensionale LAM bei hinreichender Modellgültigkeit hingegen als technologisch brauchbare Konstruktrepräsentation aufgefaßt werden. Durch die Synthese elementarer Operationen sind mit Hilfe kontextunabhängiger Schätzungen ihrer relativen Schwierigkeit, eindimensionale Aufgaben mit abschätzbarer Komplexität herstellbar.

4.2 DAS LINEARE LOGISTISCHE TESTMODELL (LLTM)

Bei den meisten Aufgabentypen wird kaum eine Möglichkeit zu finden sein, valide, nicht durch die Bedingungen der experimentellen Situation mitbedingte Daten auf der Ebene der kognitiven Operationen zu gewinnen. Die üblichen Antwortdaten auf Itemniveau enthalten nicht die erforderlichen statistischen Informationen für CML-Schätzungen der strukturellen Parameter des LAM, ebenso geht aus den Personenrandsummen der Antwortdatenmatrix nicht hervor, wieviele Operationen richtig durchgeführt wurden. Da es deshalb nicht möglich ist, die inzidentellen Parameter aus den Schätzgleichungen für die strukturellen Parameter zu eliminieren, wie dies für spezifisch objektive Operationsparameterschätzungen erforderlich wäre, ist das LAM ungeeignet, wenn ausschließlich Daten auf Itemniveau vorliegen.

SCHEIBLECHNER (1972) schlug daher erstmals vor, das RM -wie gewohnt- für Itemlösungswahrscheinlichkeiten zu formulieren, die Aufgaben charakterisierende Parameter jedoch im Sinne der oben angesprochenen Methode der Komplexitätsfaktoren als Funktion der zu ihrer Lösung erforderlichen Operationen darzustellen. Während SCHEIBLECHNER (1974) noch von einem multiplen Regressionsansatz ausgeht, um die Schwierigkeit von figuralen Aufgaben zum logischen Aussagenkalkül auf wenige logische Grundoperationen zurückzuführen, entwickelte Fischer (1972, 1973) einen Modellansatz, der Conditional-Maximum-Likelihood-Schätzungen der strukturellen, für alle Personen einer Population in gleicher Weise geltenden, Operationsparameter gestattet. Das (eindimensionale) Lineare Logistische Testmodell (LLTM) ermöglicht, bei hinreichender Geltung des Modells für einen Datensatz, allgemeine/ verallgemeinerbare Aussagen über strukturelle Parameter der experimentellen Situation trotz individueller Unterschiede. Da das LLTM bisher hauptsächlich als formales Denk- bzw. Lernmodell interpretiert wurde (vgl. SPADA, 1976) und auch die vorliegende Untersuchung von dieser Interpretation ausgehen wird, soll

das Modell nachfolgend ausschließlich unter diesem Aspekt erörtert werden. Weitere Anwendungsmöglichkeiten werden z.B. von FISCHER (1974, 1976, 1978 und FISCHER und FORMANN, 1982) vorgestellt und diskutiert.

Ein lineares logistisches "Denkmodell" formalisiert eine einfache Theorie über die Komponenten/ Operationen des kognitiven Lösungsverhaltens bei einer definierten Klasse von Problemstellungen (=kontentvaliden Aufgabenmenge). Der Grundgedanke besteht darin, die kognitive Komplexität (=Schwierigkeit) einer Aufgabe, auf die Anzahl und die Schwierigkeit der zur Aufgabenlösung erforderlichen elementaren Operationen zurückzuführen. Formal werden die Aufgabenparameter des RM als lineare Kombination gewichteter "Operationsparameter" dargestellt:

$$(10) \qquad b_g \doteq t_g = \sum_{j=1}^{m} f_{gj}\eta_j + c$$

wobei:

b(g) ... logarithmierter Schwierigkeitsparameter der Aufgabe g
t(g) ... theoretischer Schwierigkeitsparameter der Aufgabe g
f(gj) ... Häufigkeit (=Gewichtung) der Operation j in Aufgabe g
η(j) ... Schwierigkeitsparameter der elementaren Operation j (j=1 ... m)
c ... von der Normierung der Parameter b(g) abhängige Normierungskonstante

Das eindimensionale logistische Testmodell von RASCH (RM) in der Darstellung (3: S.56) geht über in:

$$(11) \qquad p\,(x_{vg} = 1 \mid v,g) = \{1 + \exp[(\sum_{j=1}^{m} f_{gj}\eta_j + c) - \theta_v)]\}^{-1}$$

Gleichung (11) spezifiziert die funktionale Abhängigkeit der Lösungswahrscheinlichkeit einer Aufgabe g von der (eindimensionalen) Fähigkeit θ der Person v zur Lösung von Aufgaben einer

homogenen Grundmenge und der Anzahl f(gj) sowie der Schwierigkeit $\eta(j)$ der hypothetisch angenommenen Lösungsoperationen. Während die Modellparameter η und θ aus den Daten geschätzt werden (siehe dazu die methodischen Überlegungen in Kp. 5.2), müssen die Operationshäufigkeiten f(gj) vor der empirischen Prüfung des Modells (s. Kp. 5.3) in Form einer hypothetischen "Aufgabenstrukturmatrix" F=(f(gj)) vorliegen. Die Matrix spezifiziert, welche elementaren Denkoperationen in welcher Häufigkeit im Lösungsprozeß der einzelnen Items der Testmenge angenommen werden können.

Führen mehrmalige Durchführungen gleichartiger Lösungsoperationen zu einer höheren Geläufigkeit dieser Operationen und damit zu einer größeren Verfügbarkeit bei der Lösung strukturgleicher oder ähnlicher Problemstellungen, so können diese, von SCHEIBLECHNER (1972) und SPADA (1976) als global oder auch operationsspezifisch gekennzeichneten, strukturellen Übungseffekte, u.U. durch Integration eines oder mehrer Transferparameter in die lineare Struktur des LLTM formal abgebildet werden. SPADA und FREY (1973) und SPADA (1973, 1976) schildern ausführlich Versuchsplanung, Datenorganisation, Parameterschätzung und Interpretation bei linear logistischen "Lernmodellen" (vgl. dazu auch Kp. 12).

4.2.1 DER OPERATIONSBEGRIFF IM LLTM

Im Vorfeld der Diskussion formaler Aufgabenbearbeitungsmodelle wurden die Begriffe "Operation" und "Komponente" relativ allgemein und unspezifisch verwendet, um analytisch und empirisch abgrenzbare Teilschritte von Aufgabenlösungsprozessen zu bezeichnen. Durch die oben dargestellte Modellierung der Aufgabenlösung ist der Operationsbegriff jedoch FORMAL präzisiert und soll nun -auch mit Blick auf seine technologische Verwendung im Rahmen der Itemkonstruktionen- entsprechend INHALTLICH bestimmt werden:

Eine "Operation" soll darin bestehen, eine sachlogisch identifizierbare Regel, die die Aufgabe/ das Problem von einem Ausgangszustand (vor der Bearbeitung) in einen Zwischen- (Teillösung) oder Endzustand (Lösung) transformiert, zu SUCHEN, zu ERKENNEN und ANZUWENDEN. Diese, aus den genannten elementaren strukturellen Denkprozessen (vgl. BÜHLER, 1907; GAGNE, 1966) zusammengesetzten "Operationen über Regeln" konstituieren die Lösungsprozesse algorithmischer Aufgaben, die als Indikatoren einer regelmengenspezifischen latenten Fähigkeit aufgefaßt werden. Die Art, Anzahl und ggfs. Verknüpfung der von ALLEN Probanden zur Aufgabenlösung durchzuführenden Operationen determiniert dabei die Komplexität der Aufgabe. Die Operation findet auf der Aufgabenseite ihre Entsprechung in der (Konstruktions-) Regel, die der Proband beim Aufgabenlösen finden UND anwenden muß; die richtige Antwort übermittelt die Informationen des Gefunden- und Angewandt- Habens bzw. die erfolgreiche Durchführung der itemschwierigkeitsbedingenden Operation.

Die Konstruktionsregeln und ihre möglichen Kombinationen beschreiben eine Aufgabengrundmenge, die ihnen entsprechenden Operationen definieren, zusammen mit der formalen Repräsentation (hier: LLTM), ein latentes Merkmal/ Konstrukt. Eine konstruierte Aufgabe ist demnach "kontentvalide", wenn sie Element dieser Itemgrundmenge ist und sie ist "konstruktvalide", wenn ihre meß-

theoretischen Eigenschaften (Homogenität, Schwierigkeit etc.) als Funktion der kognitiven Operationen über die Lösungsregeln beschreibbar sind.

Einige der o.a. definitorischen Einschränkungen gehen auf die Linearität - und Invarianzannahmen des LLTM zurück, so z.B. die Annahme gleicher (und gleich schwieriger) Operationen für alle Personen (Invarianz) oder die Annahme, daß alle denkbaren Abfolgen von Operationen, die für eine Aufgabenlösung angenommen wurden, zu ihrer Lösung führen (Linearität). Die Beschränkung auf algorithmische Lösungsprozesse und die Entsprechungsannahme von (Konstruktions-) Regel und Lösungsoperation sind hingegen weniger durch die formale Modellstruktur, als vielmehr durch die Erfordernisse einer ableitbaren Itemkonstruktionstechnologie bedingt. Konstruktrepräsentationen mit dem LLTM stellen daher zwar lediglich bereichs- oder gegenstandsspezifische "Mikro-" Theorien des Problemlösens dar, sie sind jedoch etwa im Vergleich zu manchen denkpsychologischen Ansätzen (vgl. DÖRNER, 1982) sparsam, sie können sich empirisch bewähren und damit als erklärungskräftig erweisen. Hier stellt sich nicht vordringlich die Frage, ob die angenommenen Operationen in einem kognitionspsychologischen Sinne "elementar" sind, sondern ob mit den Operationen alle möglichen Problemstellungen im vorgegebenen Aufgabenbereich erfaßt werden können und sich zu allen Operationskombinationen Aufgaben finden lassen. "Die Denkoperationen sind (...) gewissermaßen in einem "technologischen Sinne" elementar" (HEYNER, 1973, S. 213).

Das LLTM gewinnt seine Bedeutung als Konstruktionsmodell durch die Möglichkeit bei vorliegenden, stabilen Operationsparameterschätzungen und Geltung aller modellkonstituierenden Annahmen "baukastenartig" konstruktvalide, homogene Aufgaben prognostizierbarer Schwierigkeit zu konstruieren. Damit erfüllt das LLTM alle oben geforderten formalen Voraussetzungen an eine adäquate Konstruktrepräsentation. Die restriktiven Annahmen des Modells erschweren seine Anwendbarkeit bzw. inhaltliche Interpretation

(prozeßorientiertes) Denkmodell (vgl. auch die Kritik LUERS, 1981), sie sind aber Voraussetzung für seine technologische Brauchbarkeit als Konstruktionsmodell. Wenn die auf der Grundlage zuvor definierter Konstruktionsregeln erstellten Hypothesen über Lösungsoperationen die Komplexität von Aufgaben eines avisierten Typs zu erklären vermögen, reduziert sich das hier vorliegende Problem einen konstruktvaliden Itempool zu erstellen auf die regel-bare Kombination kognitiver Operationen mit -bekannten- Schwierigkeitsbeiträgen zu Aufgaben -abschätzbarer- Komplexitat.

Dieses Vorgehen setzt voraus, daß ein spezifiziertes LLTM in der gesamten definierten Personen- und Aufgabenpopulation gültig ist. "Globale" Kontrollen der Gültigkeit des Modells für einen Datensatz (d.h. für bestimmte Aufgaben die von bestimmten Probanden bearbeitet wurden) geben keine hinreichende Auskunft über seine hier vorausgesetzte Allgemeingültigkeit (s. SCHEIBLECHNER, 1975). Die Generalisierbarkeit der Modellannahmen kann erst nach Kreuzvalidierungen der referierten Ergebnisse an anderen Personen- und Aufgabenstichproben behauptet werden (vgl. Kp. 5.3 und 12). Durch die oben geforderte logische Parallelität des "Operations-" und "Regel-" Begriffs ist es dabei in der Regel sehr wohl möglich, die Konstruktionsregeln zu modifizieren oder zu revidieren, beispielsweise, wenn sich die Schwierigkeitsbeiträge bestimmter Operationen als instabil erweisen oder der gesamte angenommene Operationssatz mehr als eine latente Fähigkeit definiert.

4.2.2 ENTSTEHUNGSZUSAMMENHÄNGE VON HYPOTHESEN ÜBER KOGNITIVE LÖSUNGSOPERATIONEN

HEYNER (1973) diskutiert mehrere Verfahren, wie Hypothesen über Lösungsoperationen gewonnen werden können. Strukturanalysen der Aufgaben mißt er dabei, neben Interviews und Testverfahren, primäre Bedeutung zu. Das Ziel dieses Vorgehens besteht darin, durch sachlogische Analyse der Aufbauelemente von Items eines

bestimmten Typs, die Regeln zu identifizieren, die zur rationalen Lösung der Aufgabe beachtet und angewandt werden müssen. Dabei wird zunächst davon ausgegangen, daß Personen die Aufgabenlösung nicht erraten oder sonstwie zufällig finden. Welche Strategien zusätzlich zu den hier avisierten analytischen Verhaltensweisen zur Aufgabenlösung angewendet werden können, ist jeweils im Bezug auf das fragliche Aufgabenmaterial zu prüfen. Der Vorschlag, Strukturanalysen vorhandener Aufgabenmengen als Ausgangspunkt zu wählen, um Hypothesen über Lösungsoperationen zu entwickeln, wird von MISPELKAMP (1985a) kritisiert. Im Zusammenhang mit der Konstruktion von Sprachtests argumentiert er für eine "theoriegeleitete Aufgabenkonstruktion", die -idealiter- von einer materialunabhängigen, allgemeinen Informationsverarbeitungstheorie ausgehend, die "zu testende Fähigkeit in einer Weise (expliziert), daß diese Fähigkeit durch das geplante Arrangement der Testanforderungen in angebbarem Maße aktiviert werden kann" (MISPELKAMP, 1985b, S.2). Allein über eine genuin theoretisch orientierte, deduktive Verfahrensweise glaubt MISPELKAMP der Gefahr materialabhängiger und damit eingeschränkt gültiger Konzeptbildung entgehen zu können. Die "rationale Itemkonstruktion" sensu HABON und HORNKE (1984), die sich von MISPELKAMP's Konzeption der "theoriegeleiteten Itemkonstruktion" in erster Linie durch die Wahl des Ausgangspunkts für die Hypothesenformulierung unterscheidet, wirft MISPELKAMP (1985, S.77) vor, sie verschleiere die "Verbindung zum eigentlichen Testzweck durch die Konstruktion reiner Aufgabentheorien". Den "eigentlichen" Sinn einer Testaufgabe sieht er nicht darin "vordergründig die Fähigkeit zur Lösung (einer Testaufgabe) bei einer Person zu erfassen, sondern die Aufgabenlösung lediglich als ein Anzeichen für eine übergreifende, in anderen als dem Testkontext bedeutsame Fähigkeit des Probanden zu interpretieren."

Ein möglicher Ansatzpunkt dieser Kritik zu begegnen, ist durch die Validitätsdiskussion in Kapitel 2 gegeben. Hier wurde deutlich gemacht, daß Testdaten NICHT als Manifestationen einer

bestimmten Fähigkeit interpretiert werden sollen oder können, nur weil die Aufgabenmenge selbst wohldefiniert ist. Die vorgeschlagene Methodologie der Konstruktrepräsentationen fordert vielmehr, die Annahme eines nicht nur "im Testkontext bedeutsamen", LATENTEN Merkmals durch den Gültigkeitsnachweis eines geeigneten Meßmodells explizit zu prüfen. Die Geschichte der Testkonstruktion (vgl. z.B. ANASTASI, 1985) lehrt, daß die Chance, überhaupt solche empirischen Evidenzen aufzuweisen, in erster Linie -und vielleicht nur- für relativ eng umschriebene Gegenstands- bzw. Aufgabenbereiche mit eingeschränkten kognitiven Anforderungen gegeben ist. Ob ein so "nachgewiesenes" latentes Merkmal prognostisch "bedeutsam" im Sinne bestimmter KRITERIENBEZOGENER Validitäten ist, kann wohl nur EMPIRISCH im Hinblick auf vorhandene oder zu konstruierende Kriterienmaße beurteilt werden. Die Frage der "nomothetic span" (EMBRETSON, 1983) eines latenten Merkmals ist nachgeordnet, wird aber vermutlich gerade für solche Konstrukte positiv beantwortet werden können, deren praktische Relevanz bereits in der faktorenanalytischen Forschung deutlich wurde; wiewohl natürlich zu bemerken ist, daß diese Konstrukte nicht in der hier geforderten Weise repräsentiert sind (vgl. Kp. 2). Wenn MISPELKAMP (1985, S.78) das Fazit zieht, daß die rationale Itemkonstruktion "auf halbem Wege" stehenbleibt, so mag dies für den gegenwärtigen Stand zutreffen. Das vorgeschlagene Vorgehen ist insofern konservativ, als es an die Erfahrungen mit herkömmlichen Aufgabenformen und -mengen anknüpft, es impliziert aber nichtsdestoweniger, daß die andere Hälfte des Weges gegangen wird.

5. ÜBERLEGUNGEN ZU PARAMETERSCHÄTZUNGEN UND GÜLTIGKEITSKONTROLLEN BEIM RM UND LLTM

Die empirische Validität des Linearen Logistischen Testmodells als formale Konstruktrepräsentation erfolgt im Sinne der Modellstruktur in zwei Stufen (vgl. SPADA, 1976; NÄHRER, 1980; FISCHER und FORMANN, 1982). Im ersten Schritt ist zu prüfen, ob das RM für die beobachteten Daten hinreichend gültig ist. Die im LLTM angenommenen linearen Nebenbedingungen sind nämlich nur dann sinnvoll interpretierbar, wenn sie sich auf Lösungsprozesse von Aufgaben beziehen, die homogen im Sinne des RM sind. Erst anschließend ist der Frage nachzugehen, ob die geschätzten Aufgabenschwierigkeiten durch lineare Kombinationen der Schwierigkeit unterstellter Lösungsoperationen hinlänglich erklärbar sind. Da Aspekten der Parameterschätzung und der Modellkontrolle auch für die nachfolgend geschilderte Untersuchung zentrale Bedeutung zukommt, sollen im folgenden einige Methoden vorgestellt und im Hinblick auf ihre Anwendbarkeit im Rahmen einer umfangreichen Itempoolkonstruktion diskutiert werden.

5.1 VERFAHREN ZUR SCHÄTZUNG DER PARAMETER DES RASCH-MODELLS

Die drei verschiedenen, hier vorgestellten Verfahren zur Parameterschätzung im RM unterscheiden sich durch das Ausmaß, in dem zusätzliche mathematisch-statistische Annahmen eingehen, die NICHT konstitutiv für das RM sind. Diese Zusatzannahmen führen zu vereinfachten Schätzalgorithmen und sind gerade deshalb in hochkomplexen Anwendungsfällen von Bedeutung. WRIGHT und PANCHAPAKESAN (1969) führen Normalverteilungsannahmen für die Personen- und Itemparameter des RM ein und gelangen zu folgendem aproximativen Algorithmus (vgl. WRIGHT, 1977):

1. Gegeben seien K' Items die von N' Personen bearbeitet wurden. Da im RM keine Itemparameter schätzbar sind, wenn alle N' Perso-

nen eine Aufgabe richtig bzw. falsch beantwortet haben und analog keine Personenparameter, wenn alle K' Items richtig oder falsch beantwortet wurden, beziehen sich nachfolgende Betrachtungen auf einen um solche Pbn und Items "bereinigten" Datensatz.

2. Für die verbliebenen K Items und N Personen werden s(g), die Zahl der Personen, die Item g richtig lösen, mit g=1 bis K und n(r), die Zahl der Personen, die r Items richtig lösen, mit r=1 bis K-1, beobachtet.

3. Zu bestimmen sind:

$$(12) \qquad x_g = \ln[(N-s_g)/s_g]$$

die "Itemlogits" als logarithmiertes Verhältnis falscher zu richtiger Beantwortungen des Items g durch N Pbn,

$$(13) \qquad x = \sum_{g}^{L} x_g/K$$

der Itemlogitmittelwert,

$$(14) \qquad U = \sum_{g}^{K} (x_g - \bar{x})^2 \; / \; (K-1)$$

die Itemlogitvarianz,

$$(15) \qquad y_r = \ln[(r/(K-1))]$$

die "Personenlogits" als logarithmiertes Verhältnis richtiger und falscher Antworten auf K Items

$$(16) \qquad y = \sum_{r}^{K-1} n_r y_r/N$$

der Personenlogitmittelwert,

$$(17) \qquad V = \sum_{r}^{K-1} n_r(y_r - \bar{y})^2/(N-1)$$

die Personenlogitvarianz.

4. Wenn durch

$$X = \{(1 + U/2.89)/(1 - UV/8.35)\}^{1/2} \quad (18)$$

als "Verlängerungsfaktor" die Varianz der Itemschwierigkeiten und durch

$$Y = \{(1 + V/2.89)/(1 - UV/8.35)\}^{1/2} \quad (19)$$

die Varianz der Personenfähigkeiten berücksichtigt werden, wobei 1.7**2=2.89 bzw. 1.7**4=8.35 aus der Anpassung der Normalogive an die logistische Verteilungsfunktion resultieren, dann ist

$$b_g = Y(x_g - x) \quad (20)$$

die Schwierigkeitsschätzung für Item g

$$SE(b_g) = Y\{N/s_g(N-s_g)\}^{1/2} \quad (21)$$

der Standardfehler dieser Schätzung,

$$\theta_r = X_{Yr} \quad (22)$$

die Fähigkeitsschätzung für die Scoregruppe r und

$$SE(\theta_r) = X\{K/r(K-r)\}^{1/2} \quad (23)$$

der Standardfehler der Personenparameterschätzung.

Numerische Analysen und praktische Erfahrungen zeigen, daß diese einfachen Aproximationen, im folgenden als 'PROX' bezeichnet (s. WRIGHT und STONE, 1979) in einer Vielzahl von Situationen anwendbar und für praktische Belange ausreichend präzise sind (vgl. WRIGHT und DOUGLAS, 1975, 1977 : WILMOTT und FOWLES, 1974; RENTZ und BASHAW, 1975; ANDRICH, 1975; MEAD, 1976).

Zu ähnlich positiven Bewertungen kommen WRIGHT und DOUGLAS (1977) bezüglich eines weiteren Schätzalgorithmus, der auf dem Maximum-Likelihood Prinzip beruht und auf die oben angeführte Normalverteilungsannahmen verzichtet (WRIGHT und PANCHAPACKESAN, 1969). In diesem ebenfalls rechentechnisch relativ ökonomischen Unconditional-Maximum-Likelihood Ansatz (UML) werden Personen- und Itemparameter iterativ simultan geschätzt, d.h. die Parameterschätzungen sind wechselseitig voneinander abhängig. Die durch lokale Abhängigkeiten bedingten Inkonsistenzen können durch eine auf die jeweiligen Item- und Personenwertverteilungen bezogene Adjustierung der Item- bzw. Personenrandsummen weitgehend kompensiert werden.

Spezifisch objektive, d.h. von Personen- und Itemparameterverteilungen wechselseitig unabhängige Aufgaben- und Personenparameterschätzungen (vgl. RASCH, 1966; FISCHER, 1974) setzen die Separierbarkeit der Parameter im Schätzverfahren voraus.

Durch "Konditionierung" der Wahrscheinlichkeit aller Itemrandsummen (s(g)) auf die Personenrandsummen (=Scores, (r(v))) ist die Wahrscheinlichkeitsfunktion zur Schätzung der Itemparameter im RM nur noch von den Itemparametern b(g) selbst abhängig:

$$(24) \qquad p\{(s_g) \mid (r_v)\} = f\{(r_v), (s_g), (b_g)\}$$

Das RM ist das einzige probabilistische Testmodell, in dem die beobachteten Randsummen (s(g);r(v)) alle in den Daten über Personen- und Itemparameter enthaltenen Informationen angeben und

damit minimal erschöpfende Statistiken für diese Parameter sind (vgl. BIRNBAUM, 1968; FISCHER, 1974; ANDERSEN, 1977). Theoretisch voll befriedigend ist daher letztendlich nur der CONDITIONAL Maximum-Likelihood SCHÄTZER, der diese vorteilhaften Eigenschaften des RM nutzt, ohne die zusätzlichen Annahmen der PROX- oder UML-Schätzer einzuführen:

Gegeben sei ein Itemset mit den antilogarithmierten Parametern (ε_g), wobei $\varepsilon_g = \exp(b_g)$. Die Wahrscheinlichkeit, daß ein Pb mit dem antilogarithmierten Parameter $\xi_v = \exp(\theta_v)$ bei lokaler stochastischer Unabhängigkeit der Antworten einen Antwortvektor $(x(vg))$ liefert, beträgt

$$(25) \qquad p\{(x_{vg}) \mid \xi_v, (\varepsilon_g)\} = \prod_{g=1}^{K} [(\xi_v \varepsilon_g)^{x_{vg}}/(1 + \xi_v \varepsilon_g)] = (\xi_v^{r_v} \prod_g \varepsilon_g^{x_{vg}})/[\prod_g (1 + \xi_v \varepsilon_g)]$$

Um diese Wahrscheinlichkeit als bedingte Wahrscheinlichkeit bei gegebenem Score r(v) ausdrücken zu können, muß die Wahrscheinlichkeit bekannt sein, mit der dieser Score bei gegebenem ξ_v zu erwarten ist. Dies ist die Summe aller $\binom{K}{r}$ Ausdrücke (25), in denen sich der Vektor $(x(vg))$ zu r summiert. Diese Summe

$$(26) \qquad \gamma_r\{(b_g)\} = \sum_{\sum_g x_{vg} = r} \prod_{g=1}^{K} \varepsilon_g^{x_{vg}}$$

wird als symmetrische Grundfunktion r-ter Ordnung für die Parameter (ε_g) bezeichnet. Die Wahrscheinlichkeit genau r richtige Antworten (Scoregruppe r) bei gegebenen ξ_v und (ε_g) läßt sich nun darstellen als

$$(27) \qquad p\{r \mid \xi_v, (\varepsilon_g)\} = \sum_{\sum_g x_{vg} = r} \prod_{g=1}^{K} [(\xi_v \varepsilon_g)^{x_{vg}}]/(1 + \xi_v \varepsilon_g) = (\xi_v^{r_v} \gamma_r)/[\prod_g (1 + \xi_v \varepsilon_g)]$$

Damit beträgt die BEDINGTE Wahrscheinlichkeit, einen Vektor (x(vg)) mit den Randsummen r(v) zu erhalten, gegeben r(v):

$$(28) \qquad p\{(x_{vg}) \mid r_v, (\varepsilon_g)\}$$

$$= (p\{(x_{vg}) \mid \xi_v, (\varepsilon_g)\})/(p\{r \mid \xi_v, (\varepsilon_g)\})$$

$$= \left(\prod_{g=1}^{K} \varepsilon_g^{x_{vg}}\right)/(\gamma_r)$$

Durch die Konditionierung auf r(v) ist θ_v nicht mehr in der Gleichung enthalten. Unter der Annahme unabhängig antwortender Pbn läßt sich die bedingte Log-Likelihood der Datenmatrix ((a(vg))) nun leicht bestimmen (vgl. FISCHER, 1974; WRIGHT und DOUGLAS, 1977):

$$(29) \qquad \log \Lambda = \sum_{g=1}^{K} s_g b_g - \sum_{r=1}^{K-1} n_r \log \gamma_r$$

wobei n(r) den Umfang der Scoregruppe r bezeichnet.

Aus Gleichung (29) lassen sich Schätzalgorithmen für die Itemparameter ableiten (vgl. FISCHER, 1974; WRIGHT und DOUGLAS, 1977). Die technischen Probleme bei der Berechnung der symmetrischen Grundfunktionen γ_r und ihren ersten und ggfs. zweiten Ableitungen sind durch bestehende rekursive Beziehungen zwischen den Gleichungen durch die Differenzenmethode (bei K<40); vgl. FISCHER, 1974) oder einer Kombination aus Differenzen- und Summationsmethode (für 40<K<100); vgl. GUSTAFSSON, 1981) lösbar. Die Lösung der so berechneten Schätzgleichungen erfolgt iterativ mittels der NEWTON-RAPHSON-Methode (vgl. FISCHER, 1974; WRIGHT und DOUGLAS, 1977) oder einer "Simple-Method", die aus einem einfachen Switching zwischen linker und rechter Seite der Schätzgleichung besteht (vgl. MARTIN-LÖF, 1973; FISCHER, 1974; GUSTAFSSON, 1977, 1981).

Wegen der Symmetrie von Personen- und Itemparametern im RM gilt die Stichprobenunabhängigkeit der Parameterschätzungen prinzi-

piell auch für die Personenparameterschätzungen. Da jedoch unter realistischen Bedingungen die Anzahl der Personen sehr viel größer ist als die Anzahl der Items, ist eine CML-Schätzung der Personenparameter wegen des hohen Ordnungsgrades der Grundfunktionen unmöglich. Die mit einem gegebenen Itemsatz korrespondierende Fähigkeitsskala wird daher auch hier unter der Annahme bekannter Itemparameter geschätzt.

Unter Abwägung der Vor- und Nachteile oben angeführter Schätzverfahren erfolgen die Schwierigkeitsschätzungen für die vorliegende Untersuchung mehrstufig.

Um in einem ersten Arbeitsgang die Plausibilität der Modellannahmen für die beobachteten Datenmengen- und Teilmengen abzuschätzen, werden die Parameter mittels PROX geschätzt, wobei eine für vorliegende Belange ausreichende Ähnlichkeit der PROX- und CML-Schätzer unterstellt wird. Dieser Arbeitsgang umfaßt die Selektionen modellinkonformer Pbn und Items zur Konstruktion von 35 homogenen eindimensionalen Einzelskalen. Für die nachfolgenden Equatingprozeduren (s.u.) werden die Schwierigkeitsparameter der verbliebenen, hinreichend homogenen Items jeder Testmenge schließlich mittels CML-Verfahren geschätzt.

5.2 DIE SCHÄTZUNG DER OPERATIONSPARAMETER IM LLTM

Da für alle Personen einer definierten Population die gleiche hypothetische Aufgabenstrukturmatrix (vgl. Kp. 4.2) angenommen wird, postuliert das Modell also, daß alle Probanden die ihnen vorgelegten Aufgaben mit denselben Operationen lösen. Allgemeinpsychologische Strukturaussagen dieser Art sind möglich, da interindividuell differierende Verhaltenswahrscheinlichkeiten durch (inzidentelle) Personenparameter berücksichtigt werden. Die Operationsparameter sind dagegen, wie die Itemparameter im RM,

strukturelle Parameter, die im Sinne bedingter Inferenz ohne Berücksichtigung der Personenparameter mit CML-Verfahren schätzbar sind (vgl. FISCHER, 1972, 1973). Die bedingte Wahrscheinlichkeit der erschöpfenden Statistiken (s(j)) für die Basisparameter η_j ist nunmehr eine Funktion der erschöpfenden Statistiken x(v.), der präexperimentell gegebenen Aufgabenstruktur F und der Basisparameter η_j.

$$(30) \qquad p\ \{(s_j)\ |\ (x_{v.})\} = f\ [(x_{v.}),\ ((f_{gj})),\ (\eta_j)]$$

mit:

$$s_j = \sum_{g=1}^{K} x_{\cdot g}\, f_{gj}$$

Die CML-Schätzgleichungen werden durch Logarithmieren und partielles Differenzieren der bedingten Likelihoodfunktion nach den gesuchten Operationsparametern η_j gewonnen. FISCHER und FORMANN (1973) entwickelten einen Algorithmus, in dem das resultierende Gleichungssystem durch ein Gradientenverfahren gelöst wird (vgl. FISCHER, 1972). Die logarithmierten Parameterschätzungen η_j liegen auf einer Differenzenskala und sind bei Modellgeltung unabhängig von der Verteilung der Personenparameter in der untersuchten Stichprobe. Voraussetzung für die Unabhängigkeit der Operationsparameterschätzungen von der Stichprobe der Aufgaben, ist in erster Linie die Gültigkeit der Hypothesen, die zur Erstellung der Aufgabenstrukturmatrix geführt haben.

SPADA (1976) stellte in einer Simulationsstudie mit verschiedenen, systematisch variierten Aufgabenmatrizen fest, daß die irrtümliche Berücksichtigung irrelevanter Operationen bei einer ansonsten gültigen Aufgabenstruktur keinen Einfluß auf die Basisparameterschätzungen hat. Die Nichtberücksichtigung einer Operation, die wesentlich zur Schwierigkeit EINIGER Aufgaben beiträgt, kann jedoch die Parameterschätzungen für die berücksichtigten Operationen erheblich verfälschen. Um solche aufgabenspezifischen Effekte bei der Parameterschätzung zu vermeiden, empfiehlt sich

daher die Untersuchung umfangreicher Aufgabenmengen, in denen die Operationen vollständig kombiniert sind.

> "Für die Zuverlässigkeit der Schätzungen der Basisparameter, ist das Verhältnis 'Anzahl Aufgaben zu Anzahl Parameter' maßgebend" (SCHEIBLECHNER, 1975, S. 325).
>
> "In den bisherigen Untersuchungen wurden im Schnitt nur etwa 25 Aufgaben analysiert und auch die Größe der Vpn-Stichprobe lag bloß zwischen 200 und 300 (...). Als brauchbare Stichprobengrößen kommen (...) vielleicht 50 bis 150 Aufgaben und 500 bis 1.000 Vpn in Betracht, wobei selbstverständlich nicht alle Aufgaben allen Vpn vorgelegt werden müßten und teilweise auch parallele, d.h. strukturgleiche Aufgaben Verwendung finden können" (SPADA, 1976,S. 195f.).

Die Operationsparameter des LLTM sind bei ihrer Schätzung nicht separierbar. Insofern wird der Forderung nach spezifischer Objektivität in diesem Punkt nicht Rechnung getragen. SCHEIBLECHNER (1975, S.324) empfiehlt daher "vollständig spezifisch objektive experimentelle Versuchsanordnungen", in denen alle Faktoren, die die Lösungswahrscheinlichkeit einer Aufgabe bestimmen, vollständig kombiniert sind.

Aus statistischer Sicht muß die Itemstrukturmatrix eine weitere Bedingung erfüllen, damit die Operationsparameter des LLTM schätzbar sind. Die Matrix F=(f(gj)) besteht aus K Zeilen (#Items) und m Spalten (# Operationen). Voraussetzung für die Modellanwendung ist die lineare Unabhängigkeit von mindestens m Spalten der Matrix, d.h. es ist bereits bei Zusammenstellung der Aufgaben darauf zu achten, daß die Matrix F Rang m aufweist.

5.3 MODELLTESTS FÜR DAS RM UND DAS LLTM

Für das RM wurden mehrere Verfahren entwickelt, die Aussagen darüber ermöglichen, wie angemessen seine zentralen Annahmen für vorliegende Testdaten sind. Die gebräuchlichsten Modelltests von WRIGHT und PANCHAPAKESAN (1969), FISCHER und SCHEIBLECHNER (1979), ANDERSEN (1973) und MARTIN-LÖF (1973) unterscheiden sich einerseits darin, daß sie mehr oder minder sensitiv gegenüber Verletzungen unterschiedlicher Modellannahmen sind, andererseits aber auch wiederum durch ihre mehr oder minder haltbaren statistischen Zusatzannahmen.

Ältere Verfahren (WRIGHT und PANCHAPAKESAN, 1969; FISCHER und SCHEIBLECHNER, 1970) setzen -von denen nur subjektiv interpretierbaren graphischen Kontrollen abgesehen (vgl. FISCHER, 1974)- neben der Schätzung der Itemparameter auch die von diesen abhängigen Personenparameterschätzungen voraus. Neuere Fit-Statistiken (ANDERSEN, 1973; MARTIN-LÖF, 1973) beruhen hingegen ausschließlich auf dem CML-Ansatz und verfügen im Gegensatz zu den erstgenannten Verfahren über asymptotisch bekannte Eigenschaften. Sie thematisieren die für das RM zentrale Annahme spezifisch objektiver Vergleiche von Personen und Items. Item- und Personenparameterschätzungen aus partitionierten Pbn- bzw. Itemstichproben sind bei Gültigkeit dieser Implikationen bis auf Schätzfehler invariant. Bei überzufälligen Abweichungen der Schätzer aus unabhängigen, exhaustiven Gruppierungen eines Datensatzes nach Probanden- oder Itemmerkmalen, sind mehr als ein Parameter zur Beschreibung einer jeden Itemschwierigkeit bzw. jeder Pbn-Fähigkeit erforderlich. GUSTAFSSON (1980) unterscheidet hier, je nach dem ob die Invarianz von Item- oder Personenparameterschätzungen untersucht wird, zwischen ICC- (item characteristic curve) Tests und PCC- (person characteristic curve) Tests.

Zu den ICC-Tests gehört der Likelihoodquotiententest von ANDERSEN (1973b), der sehr eng auf die CML-Schätzmethode bezogen und daher

theoretisch befriedigend ist. Die Teststatistik stellt das Maximum der (Log-)Likelihood der Gesamtdaten unter Bedingung der aus ihnen geschätzten Itemparameter L(t) den summierten Maxima der (Log-)Likelihoods der nach Pbn-Charakteristika geteilten Daten unter der Bedingung der aus diesen g disjunkten Teilmengen geschätzten Parametern L (j) (j = 1,...,g) einander gegenüber:

$$(31) \qquad \log(\lambda) = L_t - \sum_{j=1}^{g} L_j$$

ANDERSEN (1973b) konnte zeigen, daß die Prüfgröße $-2\log(\lambda)$ für $n \to \infty$ mit df=(g-1)(K-1) asymptotisch CHI**2-verteilt ist.

Je nach dem, welche Kriterien zur Gruppierung der Personen herangezogen werden, ist der Test sensitiv gegenüber verschiedenen Verletzungen der Modellannahmen.

Wird nach Scoregruppen partitioniert, ist der Test sensitiv gegenüber Abweichungen in der Steigung der ICC und somit geeignet, das RM gegen die Annahme zu testen, daß evtl. das BIRNBAUM-Modell (vgl. LORD und NOVICK, 1968) oder ein Modell mit noch mehr Parametern zur angemessenen Beschreibung der Daten erforderlich ist.

Meistens wird die Invarianz der Schätzungen aus hinsichtlich Rohscore oder externen Kriterien, wie Alter, Geschlecht, Schulbildung etc. dichotomisierten Datensätzen überprüft. Bei Unterteilungen nach externen Kriterien thematisiert der Modelltest in erster Linie die Frage, ob der Test gegenüber dem jeweiligen Kriterium "biased" ist. Sind Kriterium und Rohwerte unkorreliert, handelt es sich darüberhinaus auch hier um einen "slope-" Test der itemcharakteristischen Kurven des RM.

Spät wurde erkannt, daß insbesondere die Unterteilung nach Rohwerten unter bestimmten Bedingungen, so z.B. wenn zwei Dimensionen durch dieselbe Anzahl von Items vertreten sind und Item- wie

zugehörige Personenparameterverteilung übereinstimmen, ungeeignet ist, selbst massive Inhomogenitäten aufzudecken (STELZL, 1979; VAN DEN WOLLENBERG, 1979).

VAN DEN WOLLENBERG (1979, S.107ff.) schlägt daher vor, zusätzlich zu den üblichen Modellkontrollen, Teile des Antwortvektors oder einzelner Items selbst als Teilungskriterien zu verwenden, wobei die jeweils zur Teilung herangezogenen Items als externe Information betrachtet und von der anschließenden Analyse auszuschließen sind. FORMANN (1981), der die höhere Trennschärfe dieses Verfahrens gegenüber Teilungen nach Rohscores anhand inhomogener, simulierter Datensätze aufzeigen konnte, motiviert die Verwendung von Items als Teilungskriterien folgendermaßen:

> "Erfaßt Item i die latente Dimension x, werden jene Personen, welche Item i gelöst haben, im allgemeinen bezüglich dieser Dimensionen leistungsfähiger sein als jene Personen, welche Item i nicht gelöst haben. Die Teilung der Gesamtstichprobe in zwei Untergruppen entsprechen dem Kriterium 'Item i gelöst' versus 'Item i nicht gelöst' wird daher weitgehend einer Teilung hinsichtlich der Leistungsfähigkeit im Bezug auf Dimension x gleichkommen" (S. 554).

Ein weiterer ICC-Slope-Test wurde von MARTIN-LÖF (1973; s. auch GUSTAFSSON, 1980) entwickelt. Er ist asymptotisch äquivalent dem Likelihoodquotiententest von ANDERSEN, unterscheidet sich im Aufbau aber wesentlich von diesem. Hier sei nur darauf hingewiesen, daß es sich um eine CHI**2-Statistik handelt, der die Abweichungen zwischen modellmäßig erwarteten und beobachteten Lösungshäufigkeiten eines jeden Items in jeder Scoregruppe zugrundeliegen. Im Gegensatz zum ANDERSEN-Test werden die Itemparameter nur aus der gesamten Datenmatrix geschätzt. Dennoch ist die Berechnung der Prüfstatistik gerade für längere Tests rechentechnisch sehr aufwendig, da so viele Kovarianzmatrizen invertiert werden müssen, wie Scoregruppen vorhanden sind. Außerdem

sind größere Stichproben erforderlich als beim ANDERSEN-Test, damit die asymptotischen Eigenschaften der Statistik greifen.

Die angesprochenen Modellkontrollen sind ICC-Tests, da sie die Invarianz der Itemparameter für verschiedene Personengruppen prüfen. Es läßt sich aber zeigen, daß diese Tests nicht gegenüber allen Verletzungen der Modellannahmen des RM sensitiv sind. So ist beispielsweise durchaus denkbar, daß die ICC "parallel", die personencharakteristischen Kurven (PCC) jedoch nicht "parallel" sind (vgl. LUMSDEN, 1978). Um die Annahme zu prüfen, ob zusätzliche "Person-Sensivity-Parameter" erforderlich sind, könnte die Invarianz von Personenparameterschätzungen für unterschiedliche Itemteilmengen geprüft werden. Dem steht allerdings gegenüber, daß sehr viele Parameter geschätzt werden müßten und praktisch keine Personenparameterschätzung auf der Grundlage bedingter Inferenz möglich ist.

MARTIN-LÖF (1973) zeigte, daß auch hier auf die Schätzung der Personenparameter verzichtet werden kann. Er formulierte ein lediglich auf CML-Itemparameterschätzung beruhenden Likelihood-quotiententest für die Hypothese, daß zwei disjunkte Gruppen von Items k(1) und k(2) dieselbe Fähigkeit messen.

Wird die Itemmenge beispielsweise bezüglich relativer Lösungshäufigkeiten geteilt, so ist er ein PCC-Slope-Test, der die Frage thematisiert, ob zusätzliche personensensitive Parameter, wie etwa im LUMSDEN-Modell (1978), erforderlich sind. Eine Dichotomisierung nach hypothetischen, inhaltlichen Kriterien, wie sie hier durch die Konstruktionsregeln bzw. Operationen gegeben sind, prüft direkt die Annahme, daß die verschiedenen Itemgruppen ein und dasselbe Merkmal erfassen.

Das zentrale Problem, ob überhaupt und welche der oben dargestellten, theoretisch befriedigenden Tests im Zusammenhang mit einer umfangreichen Itempoolkonstruktion angewendet werden

sollten, um den Fit des RM zu überprüfen, muß auch und besonders unter Praktikabilitätsaspekten betrachtet werden. Um die ANDERSEN- beziehungsweise MARTIN-LÖF-ICC- und PCC-Tests als Slope-Tests sinnvoll interpretieren zu können, müssen die Daten in äußerst umfangreichen Pbn-Stichproben erhoben werden, wobei die erforderlichen Stichprobengrößen auch von den Testlängen abhängen.

So sind beispielsweise im ANDERSEN ICC-Slope-Test die Itemparameter nicht nur für die in der Regel mit ausreichenden Datenpunkten besetzten "mittleren", sondern auch für die meist schwach besetzten "extremen" Scoregruppen zu schätzen. Letztere müssen zusammengefaßt werden, wenn die Verläufe der bedingten Likelihoodfunktionen die Bestimmung von Maxima nicht sinnvoll erscheinen lassen. Dies führt aber dazu, daß der Test -im strengen Sinne- nicht mehr als Slope-Test interpretiert werden kann.

Ähnliche Einschränkungen sind auch dann zu erwarten, wenn Items als Teilungskriterium für die Pbn-Stichprobe verwendet werden (VAN DEN WOLLENBERG, 1979; FORMANN, 1981). Wenn z.B. Items mit einer relativen Lösungshäufigkeit $p<.20$ oder $p>.80$ als Teilungskriterien herangezogen werden, liegt der Gesamtstichprobenbedarf bei $N>500$, damit jede der beiden Gruppen mit gleichen Itembeantwortungen mit minimal $n=100$ Pbn besetzt ist. Für die Evaluation eines einzelnen Tests ist ein solcher Stichprobenumfang durchaus realistisch und im Rahmen von RASCH-Analysen inzwischen nicht unüblich. Bedenkt man allerdings, daß unter gleichen Randbedingungen im Rahmen der vorliegenden Itempoolkonstruktion für 35 Tests $N>17.500$ anzustreben wäre, wird die Anwendungsproblematik dieses Modelltests deutlich. Der ICC-Slope-Test von MATRIN-LÖF wird wegen seines rechentechnischen Aufwandes selbst für Standarduntersuchungssituationen nur bedingt empfohlen (vgl. GUSTAFSSON, 1981, S.9). In komplexen Anwendungsfällen sprechen beschränkte Rechnerkapazitäten und damit verbundene Kosten-Nutzen-Überlegungen gegen seine Verwendung.

Ein weiterer Aspekt ist prinzipieller Natur und betrifft die Nützlichkeit dieser Verfahren für die Evaluation einer rationalen Itemkonstruktion überhaupt.

Bei den dargestellten ICC- und PCC-Tests handelt es sich um GLOBALE Modellkontrollen, die zwar Aussagen darüber zulassen, ob eine Itemmenge homogen im Sinne des Modells ist, darüberhinaus aber keinerlei Informationen über mögliche Gründe für einen schwachen Modellfit liefern.

Gerade im Hinblick auf die hier angestrebte Validierung einer Itemkonstruktions- und Bearbeitungstheorie scheinen Modellkontrollen ökonomischer und zielführender, die, wenn auch wegen ihrer theoretischen Mängel, vorsichtig zu interpretierende Gültigkeitsaussagen auf Itemebene erlauben.

Solche Teststatistiken wurden von FISCHER und SCHEIBLECHNER (1970) und WRIGHT und PANCHAPAKESAN (1969; vgl. auch WRIGHT, MEAD und DRABA, 1976; MEAD, 1976) entwickelt. Sie setzen die Schätzung der Personenparameter des RM voraus.

FISCHER und SCHEIBLECHNER (1970) führen eine Teststatistik ein, die ebenfalls die zentrale Annahme der (Teil-) Stichprobenunabhängigkeit der CML-Itemparameterschätzungen thematisiert. Der Test basiert auf einem Vergleich der logarithmierten Schwierigkeitsschätzungen aus zwei disjunkten Datenteilmengen (1) und (2). Über die Informationsfunktion des Test in Bezug auf die geschätzten Parameter b(1) und b(2) ist folgende asymptotisch standardnormalverteilte Teststatistik u formulierbar:

$$(32) \qquad u_g = (b_g(1) - b_g(2))/(s_g(1)^2 + s_g(2)^2)^{1/2}$$

wobei s(1) die Standardabweichung des Schätzers b in der Teilstichprobe (1) bezeichnet. Nach FISCHER und SCHEIBLECHNER (1970, S33) ergibt sich aus der Summe von (32) über alle Items eine

Globalstatistik für die Gesamtanpassung des Tests:

(33) $$S = \sum_{g} u_g^2$$

Unter der Annahme unabhängiger Itemparameterschätzungen folgt S einer CHI**2-Verteilung mit k-1 Freiheitsgraden. Da diese Annahme jedoch nicht haltbar ist, wie FISCHER (1974) selbst anmerkt, und VAN DEN WOLLENBERG (1979) zeigen konnte, daß der Erwartungswert von S systematisch von der Anzahl der Freiheitsgrade abweicht, ist es nicht angezeigt, S in der vorliegenden Untersuchung im Hinblick auf Selektionsentscheidungen zu interpretieren. Hier soll vielmehr versucht werden, Variabilitäten in den Itemparameterschätzungen bezüglich verschiedener Partitionierungskriterien ohne globale, auf den Test bezogene Anpassungsstatistiken aufzudecken.

Letztendlich sind aber auch die erschöpfenden Statistiken, aus denen die Item- (und Personen-) Parameter geschätzt werden, Summen von Einzelbeobachtungen, die jede für sich im Rahmen des Modells eine bestimmbare Plausibilität besitzen. WRIGHT und STONE (1979) wählen daher, aufbauend auf WRIGHT und PANCHAPAKESAN (1969) und MEAD (1975), konsequenterweise die einzelne Pbn-Antwort zum Ausgangspunkt ihrer Überlegungen zum Modellfit.

Bei gegebenem Aufgabenparameter b(g) und Personenparameter $\Theta(v)$ schätzt die Funktionsgleichung des RM die Wahrscheinlichkeit Π_{vg}, mit der eine Person v eine richtige (x(vg)=1) bzw. falsche (x(vg)=0) Antwort auf Item g liefert. Wenn man in diesem Sinne p(vg) als Schätzung für die erwartete Ausprägung von x(vg) interpretiert und

(34) $$p_{vg} (1 - p_{vg})$$

als Schätzung für deren erwartete Varianz, so ist mit

$$(35) \qquad z_{vg} = (x_{vg} - p_{vg})/[p_{vg}\ (1 - p_{vg})]^{1/2}$$

für jede Realisation von x(vg) ein Residuum bestimmbar, das die geschätzte binomiale Standardabweichung solcher Beobachtungen darstellt. In dem Ausmaß, wie das Modell den Daten angemessen ist, werden die geschätzten Residuen z(vg) annähernd standardnormalverteilt sein (vgl. WRIGHT und STONE, 1979, S.69ff.):

$$(36) \qquad z_{vg} \sim N(0,1)$$

Die systematische Untersuchung der Residuen erlaubt es, sehr datennah und eher deskriptive als schlußfolgernde Aussagen darüber zu machen, wie plausibel es ist anzunehmen, eine Person antworte konsistent im Sinne des Modells und Testexperiments oder ein Item erfasse nicht mehr als die zugrundeliegende latente Fähigkeit. Treten bei einer Aufgabe zahlreiche in diesem Sinne "erwartungswidrige" Antworten auf, stellt sich die Frage, ob das Item für Messungen der fraglichen Pbn-Fähigkeit geeignet ist. Wie bei allen bisher dargestellten Gültigkeitskontrollen wird die Inhomogenität des Items als eine mögliche Quelle mangelnder Gültigkeit betrachtet. Diese "einseitige" Sichtweise wird aber durch das Meßmodell selbst keineswegs nahegelegt, da es Items und Probanden völlig symmetrisch "behandelt". Sie ist vielmehr notwendige Folge der Dissoziation traditioneller Testkonstruktion und -anwendung. Der Konstrukteur eines herkömmlichen Papier- und Bleistift- Tests hat eine Zielpopulation vor Augen, die er in der Regel über einige wenige charakteristische Merkmale zu beschreiben versucht. Aus empirischen Evidenzen für die Gültigkeit des gewählten Testmodells in einer Stichprobe, auf die die definitorischen Populationsmerkmale zutreffen, generalisiert er die Annahmen der Modellgültigkeit auf die gesamte Pbn-Population (vgl. etwa die Überlegung zur Konstruktion des Wiener Matrizentests von FORMANN und PISWANGER (1979)). Dabei handelt es sich

aber um eine idealisierte Sichtweise, die davon ausgehen muß, daß die Itemmenge bei allen zukünftig zu messenen Probanden nur das fragliche Merkmal erfaßt. Sie ist bei Anwendung des RASCH-Modells aber unzulässig, da die Bedeutungsinvarianz der Items nur für Pbn behauptet werden kann, die nachweislich modellkonformes Antwortverhalten zeigen. D.h.: Die Frage der Meßbarkeit eines Pb mit einem RASCH-homogenen Itemsatz ist nur empirisch lösbar und muß für jeden getesteten Pb post hoc entschieden werden.

Diese Sichtweise mag dem traditionell denkenden Testanwender/ Diagnostiker fremd erscheinen, sie wird aber vielleicht durch die konsequente Fortführung eines häufig zitierten Beispiels plausibel: als wesentlicher Vorteil des RM gegenüber anderen Testmodellen wird häufig darin gesehen, daß ein modellkonformer Test spezifisch objektive, d.h. vom Meßinstrument unabhängige Vergleiche zwischen Pbn hinsichtlich eines Merkmals erlaubt. Dies wird hier und da (vgl. z.B. WRIGHT und STONE, 1979) am Beispiel einer Meßlatte veranschaulicht, die Größenvergleiche zwischen Personen ermöglicht, unabhängig davon, wer gemessen wird. Man stelle sich aber vor, eine Person x verhalte sich bei dem Versuch, ihre Größe zu ermitteln, anweisungswidrig, indem sie beispielsweise ständig mit dem Kopf wackelt oder Kniebeugen vollführt. Niemand würde in einer solchen Situation die Meßlatte zum ungeeigneten Instrument für Größenmessungen erklären. Vielmehr würde man das Scheitern der Messung dem unkooperativen Verhalten der Person x zuschreiben.

Auf die psychologische Testsituation übertragen bedeutet dies, daß die "Meßbarkeit" eines Pb fraglich ist, wenn sich der Pb der experimentellen Bedingung entzieht. Personenspezifisches Antwortverhalten, das der Modellerwartung widerspricht, drückt sich in einer Häufung großer Residuen z(vg) für diese Person aus. Summiert man nun die quadrierten Residuen für einen Pb über alle Items, so erhält man eine annähernd CHI**2-verteilte Teststatistik v, die nach WRIGHT und STONE (1979) eine statistische

Bewertung des Personenfit erlaubt:

$$(37) \qquad V_v = \sum_g z_{vg}^2$$

Relativiert man v auf die Anzahl der Freiheitsgrade f=k-1, resultiert eine approximativ standardnormalverteilte Teststatistik t:

$$(38) \qquad y_v = \sum_g z_{vg}^2/f$$

$$(39) \qquad t_v = [\ln(y_v) + y_v - 1](f/8)^{1/2} \sim N(0,1)$$

Wird beispielsweise in einer computerunterstützten Testsituation der Antwortvektor eines Pb dokumentiert, ist mit Hilfe der Personenmisfitstatistik t(v) abschätzbar, wie plausibel es ist, die Populationszugehörigkeit des Pb anzunehmen. Bei der Konstruktion eines Itempools können über t(v) "auffällige" Pbn selegiert werden, da die "Informationsbeiträge" ihrer Antwortvektoren zur Schätzung der hier vorrangig interessierenden Itemparameter mit einiger Wahrscheinlichkeit gering sind.

Analog zur Personenmisfitstatistik t(v) wird von WRIGHT und STONE (1979, S.77ff.) eine Itemmisfitstatistik t(g) vorgeschlagen:

$$(40) \qquad y_g = \sum_v z_{vg}^2/(N-1)$$

$$(41) \qquad t_g = [\ln(y_g) + y_g - 1](N-1/8)^{1/2} \sim N(0,1)$$

Zu den statistischen Eigenschaften der t(g)- bzw. t(v)- Statistik schreiben WRIGHT und STONE (1979, S.168ff.):

> "The fit statistik t is distributed more or less normally but with wider tails. In our practical experience a popular rejection level of about two is unnecessarily conservative. The general guide-

lines we currently use for interpreting t as a signal for misfit are:

If t<3, we accept the measurement of the person as probably valid.

If 3<t<5, we make a careful examination of the response pattern in order to identify and consider possible sources of misfit.

If t>5, we reject the measure, as it stands..."

Selektionsentscheidungen sollten sich generell nicht schematisch an kritischen Teststatistiken orientieren (vgl. HATTIE, 1985). Im vorliegenden Fall können verfügbare Informationen über die Einzelbeiträge (z_{vg}^2) der Summenstatistiken mitberücksichtigt werden. So scheint es beispielsweise nicht angebracht, ein Item allein wegen seines auffälligen t(g)-Wertes zu seligieren, wenn die zugehörige, ansonsten homogene Residuenverteilung wenige, durch vereinzelte erwartungswidrige Pbn-Antworten bedingte Extremwerte enthält, die allein durch ihren überproportionalen Beitrag die Summenstatistik "aufblähen". Es wird daher ergänzend vorgeschlagen, nur dann über den t-Wert auffällige (t(g)>5) Aufgaben zu seligieren, wenn mindestens 5% aller Residuen z_{vg}^2 größer sind als ein kritischer Wert (hier z(vg)**2>5 .

Bei der Konstruktion homogener Skalen mit dem RASCH-Modell werden bei mangelnder Modellgeltung überlicherweise Items (oder Pbn) selegiert, selbst wenn sie zweifelsfrei einer zuvor definierten Grundmenge bzw. Population zugeordnet werden können. Damit wird jedoch die inhaltliche Grundmengen-/ Populationsdefinition negiert und der Gültigkeitsbereich sämtlicher Aussagen über Personen und Items auf die verbleibenden modellverträglichen Items (und Personen) reduziert. Es ist dann nicht mehr möglich anzugeben, welche Aufgaben die den Latent Trait inhaltlich definierende Grundmenge konstituieren bzw. welche Items potentiell als homogene Aufgaben in Frage kommen.

Die bloße technische Selektion von Aufgaben erbringt keinerlei Information über Prinzipien zur Konstruktion homogener konstruktvalider Testaufgaben. GITTLER (1983) führt zudem häufig zu beobachtende alternierende Inhomogenitäten in Kreuzvalidierungen bereits "bereinigter" Itemmengen auf ausschließlich technische Selektionsstrategien zurück. HILKE (1980, S.352) fordert:

> "daß der inhaltlichen Definition der Population das Primat zukommt. In diesem Fall verbietet sich bei Verwerfung des Modells die Elimination von Items oder Personen von selbst. Wir sind vielmehr gehalten, alle unsere theoretischen Annahmen -einschließlich der Festlegung der Population- zu überprüfen und ggfs. zu modifizieren bzw. zu spezifizieren, sowie einen neuen Falsifikationsversuch zu wagen."

Modellanpassung durch Elimination inhomogener Items ließe sich dann nur mit dem Informationsgewinn begründen, der sich z.B. aus einer Analyse struktureller Merkmale selegierter Aufgaben für die (Re-)Definition des Konstruktionsansatzes ergeben könnte. Es sollte daher stets versucht werden, aus Strukturanalysen "technisch" auffälliger Items post hoc verallgemeinerbare, d.h. keine singulär geltenden (!) inhaltlichen Selektionsgründe in den Regel- bzw. Operationsbegriffen des Konstruktionsansatzes zu formulieren. Die Validität des ensprechend eingeschränkten Konstruktionsrationals ist dann in weiteren Konstruktionsexperimenten empirisch zu prüfen.

Bei hinreichender Geltung des RM ist dann zu prüfen, ob die Nullhypothese der linearen Nebenbedingungen beibehalten werden kann, d.h. gemäß (10; S.74) wird kontrolliert, ob die Itemparameterschätzungen b(g) hinreichend genau durch die im LLTM formulierte additive Kombination t(g) gewichteter Operationsparameter darstellbar sind.

Vom trivialen Fall m=k abgesehen, wird die unabhängige Variier-

barkeit der Itemparameter bei Modellgeltung durch die angenommenen linearen Restriktionen beträchtlich eingeschränkt. Die bei der Parameterschätzung maximierte bedingte Wahrscheinlichkeit (30; S.88) wächst aber mit der Zahl frei variierender Parameter. Um nun eine Aussage über die Angemessenheit der Nebenbedingungen treffen zu können, liegt es nahe, folgenden Likelihoodquotienten zu betrachten:

$$(42) \qquad \lambda = L_0[(x_{\cdot g})|(x_{v\cdot}),(t_g)]/L_1[(x_{\cdot g})|(x_{v\cdot}),(b_g)]$$

Gilt das Modell, so sollte L(0) nur wenig kleiner als L(1) bzw. der Quotient nur geringfügig kleiner als 1 sein. Die Prüfung der statistischen Signifikanz der Abweichung erfolgt über die Prüfgröße $Z = -2 \ln\lambda$ des Conditional-Likelihood-Ratio Tests (CLR) (vgl. ANDERSEN, 1973; FISCHER, 1974). Die Prüfgröße Z ist für $n \to \infty$ asymptotisch CHI**2-verteilt mit (k-m) Freiheitsgraden, wobei k und m die Anzahl unabhängig geschätzter Item- bzw. Operationsparameter bezeichnen.

Bisherige Anwendungen haben gezeigt, daß CLR-Tests streng sind und die Prüfgröße Z schon bei betragsmäßig geringfügigen Abweichungen zwischen den Itemparametern b(g) und den theoretischen Aufgabenschwierigkeiten t(g) signifikante Werte annimmt (vgl. NÄHRER, 1977). Man wird also für jeden Anwendungsfall zu prüfen haben, ob der Test überhaupt eingesetzt werden soll bzw. welche Ungenauigkeiten im Hinblick auf PRAKTISCHE ZWECKE tolerierbar sind.

Bei mangelnder Modellkonformität empfiehlt es sich jedoch stets, mittels graphischer Kontrollen den Ursachen einzelner Abweichungen nachzugehen. Zu diesem Zweck werden die Punkte P(t(g);b(g)) in ein Koordinatensystem eingetragen und ihre Abweichungen von der durch den Ursprung verlaufenden 45 Grad-Geraden beurteilt. Ist die Anpassung perfekt, so liegen alle Punkte auf dieser Geraden.

Um die Situations-, Personen- und Aufgabenstichprobenunabhängigkeit der Operationsparameterschätzungen zu beurteilen, schlägt NÄHRER (1977, 1980) detailierte Modellkontrollen in Datensegmenten sowie Kreuzvalidierungen in anderen Aufgaben- und Personenstichproben vor. Einige dieser Verfahren werden in der vorliegenden Untersuchung angewendet und dort näher erörtert.

6. KONKRETISIERUNG DER PROBLEMSTELLUNG

In den folgenden Kapiteln wird der Versuch berichtet, eine Konstruktrepräsentation für figurale Denkaufgaben zu entwickeln und zu erproben. Bei dem avisierten Aufgabentyp handelt es sich um sog. Matrizenaufgaben, zu deren Lösung Regeln beachtet und angewendet werden müssen.

Untersuchungen zu diesem Aufgabentyp haben eine lange Tradition, die bis in die Anfänge psychologischer Intelligenztheoriebildung (SPEARMAN, 1904) zurückreicht. Es würde an dieser Stelle zu weit führen, die zahlreichen faktorenanalytischen Untersuchungen allein in Bezug auf die bekannten "Progressive Matrices" von RAVEN (1956, 1962) darzustellen. Einen recht umfassenden Überblick hierzu gibt PUTZ-OSTERLOH (1981). Sie kommt zu dem Schluß, daß diese Untersuchungen mit den gleichen Schwächen behaftet sind, wie nahezu alle anderen, die versuchten oder versuchen, mit Hilfe explanatorischer Faktorenanalysen die Validität von Intelligenztests zu bestimmen (vgl. dazu die Kritik in Kp. 3). ERTEL (1966, zit. nach PUTZ-OSTERLOH, 1981) faßt das Ergebnis dieser Bemühungen treffend zusammen:

> "Die ungeheueren empirischen Bemühungen zum Problem der Intelligenz in den letzten Jahrzehnten haben zwar eine Vielzahl detaillierter Befunde gebracht. Doch gibt es bis heute keine gesicherte theoretische Fundierung der Intelligenz (...).

> Auch die zu Beginn der Entwicklung so viel versprechende Methode der Faktorenanalyse ist trotz mancher Ansätze ohne allgemein verbindlichen Ertrag geblieben" (S.152).

Als Alternative zu faktorenanalytischen Validitätsbestimmungen für Intelligenztests werden gerade in jüngerer Zeit denkpsychologische Untersuchungsmethoden vorgeschlagen. PUTZ-OSTERLOH (1981) versucht auf diese Weise, das Problemlösungsverhalten beim Bearbeiten der "Advanced Progressive Matrices" (APM; RAVEN, 1961) zu analysieren. Sie beschreibt ihr Vorgehen folgendermaßen:

> "Die Aufgaben des Intelligenztests wurden in ihrer Struktur beschrieben und klassifiziert. Aus dieser theoretischen Strukturbeschreibung wurden Hypothesen darüber abgeleitet, in welchen Problemlösungsschritten sich richtig-lösende (R-Löser) von falsch-lösenden Probanden (F-Löser) unterscheiden müßten" (PUTZ-OSTERLOH, 1981, S.208).

Wie der kurzen Beschreibung zu entnehmen ist, entspricht der Entstehungszusammenhang der Hypothesen über Lösungsoperationen bzw. Denkprozesse dem hier vorgeschlagenen Verfahren, nämlich von Strukturanalysen des betreffenden Aufgabenmaterials auszugehen. Ziel und Methode des weiteren Vorgehens unterscheiden sich jedoch wesentlich von denen einer formalen Konstruktrepräsentation. Mit Hilfe von Blickbewegungsregistrierungen und anhand von Daten des "Lauten Denkens" werden in der Regel einzelne Aspekte individuellen Lösungsverhaltens weniger Probanden bei einer geringen Anzahl von Aufgaben untersucht. Strukturelle, allgemeinpsychologische Aspekte sind dabei von eher geringer Bedeutung, weshalb die Ergebnisse in der Regel nur für die untersuchten Probanden und Aufgaben Gültigkeit beanspruchen können. Nichtsdestoweniger können diese Ergebnisse aber bei der Generierung von Hypothesen im Rahmen einer Konstruktrepräsentation sinnvoll verwendet werden.

In den siebziger Jahren wurde an der Universität Wien ein formales Lösungsmodell für Matrizenaufgaben auf der Basis des LLTM entwickelt. Dieser Ansatz soll hier nun um einige besonders relevant erscheinende Ergebnisse denkpsychologischer Untersuchungen, besonders der bereits zitierten Studie PUTZ-OSTERLOH's erweitert werden. Ziel dieses Vorgehens ist, durch VOLLSTÄNDIGE Kombination aller dann definierten Aufgabenstrukturelemente, ein Universum von Matrizenitems zu erstellen, das einerseits den oben postulierten validitätskritischen und testtheoretischen Kriterien genügt und damit auch für das computerunterstützte adaptive Testen geeignet ist und zum anderen die Möglichkeit eröffnet, das LLTM mit den in Kapitel 5 vorgestellten Methoden als zugrundeliegende formale Konstruktrepräsentation zu validieren.

7. DIE MODELLIERUNG KOGNITIVER LÖSUNGSOPERATIONEN BEI MATRIZENAUFGABEN

7.1 DIE UNTERSUCHUNGEN VON FORMANN UND NÄHRER

Mit dem LLTM wurden schon mehrfach kognitive Lösungsoperationen struktureller Aufgaben beschrieben (vgl. FISCHER und FORMANN (1982), die einen Überblick zu bisherigen Anwendungen des LLTM geben). Denkaufgaben vom Typ der "Progressive Matrices" (RAVEN, 1956) wurden erstmalig von FORMANN (1973: s. auch FORMANN und PISWANGER, 1979) analysiert. Zu diesem Zweck konstruierte FORMANN systematisch neue Aufgaben, indem er von zuvor definierten Konstruktionsregeln ausging. Wie bei diesem Aufgabentyp üblich, ist dabei die fehlende Figur im Aufgabenschema (vgl. Abb.4) durch logische Beziehungen zwischen den aus Einzelkomponenten aufgebauten 8 Figuren des Schemas eindeutig festgelegt. FORMANN beschreibt die Einzelkomponenten durch ihre MATERIALEIGENSCHAFTEN (ME): FORM (Fo), MUSTER (M), ANZAHL (A) und RÄUML. ANORDNUNG (R).

Die Aufbaugesetzmäßigkeit des Schemas wird durch REGELN (RE) hergestellt, die über die relevanten Materialeigenschaften einzelner Komponenten definiert sind und die von den Probanden gefunden und zur Lösungsfindung angewendet werden müssen. Es sind dies

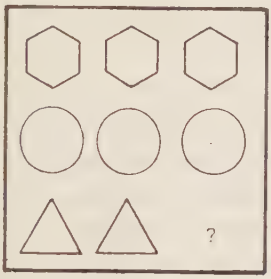
a) FORTSETZEN (F) desselben graphischen Repräsentanten einer Materialeigenschaft je Zeile und/oder Spalte des Schemas,

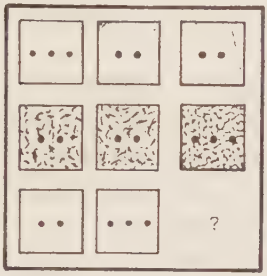
b) VARIIEREN (V) der Reihenfolge von drei Repräsentanten einer Materialeigenschaft je Zeile/Spalte,

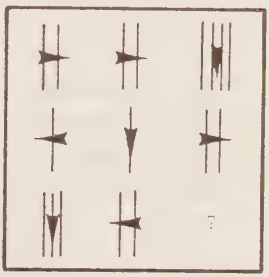
c) ÜBERLAGERN (Ü) von Repräsentanten einer ME in den Zeilen/Spalten des Schemas.

Ein weiterer "Faktor" mit den Stufen WAAGERECHT (w), SENKRECHT (s) und WAAGERECHT/SENKRECHT (w/s) definiert die jeweilige RICHTUNG, in der die REGEL innerhalb des Schemas anzuwenden ist. Die Konstruktionsregeln definieren, zusammen mit der allgemeinen Aufgabenform und einer bei FORMANN noch nicht spezifizierten Menge graphischer Repräsentanten ein Universum von Matrizenaufgaben.

Auf der Grundlage der Strukturbeschreibung von 42 entsprechend konstruierten Items formulierte FORMANN mit dem LLTM eine sparsame Theorie kognitiver Lösungsoperationen, die davon ausgeht, daß Probanden die Items lösen, indem sie herausfinden, in welcher Richtung des Schemas welche Regeln für welche Materialeigenschaften gelten.







KOMPONENTE 1			
Material:	Form	Muster	Anzahl
Regel :	Fortsetzen	Überlagern	Überlagern
Richtung:	waagrecht	waagrecht	waagrecht
KOMPONENTE 2			
Material:	--	Anzahl	Räuml.Anordn.
Regel :	--	Variieren	Variieren
Richtung:	--	waagr./senkr.	senkrecht

Abbildung 4:
Beispiele für regelgeleitet konstruierte Itemschemata

In einer empirischen Untersuchung der 42 so konstruierten Aufgaben mußte FORMANN, entgegen der Erwartung, 9 Items selegieren, die sich nicht als homogen im Sinne des RM erwiesen. Für die verbleibenden Aufgaben wurden zahlreiche konkurrierende Hypothesen zu Art, Anzahl und sogar Interaktion der am Lösungsprozeß beteiligten kognitiven Operationen mit dem LLTM geprüft. Unter der Annahme, daß die drei Faktoren 'Materialeigenschaften', 'Regeln' und 'Richtung der Regelgeltung' unabhängig voneinander schwierigkeitserzeugend wirken, konnten aus den Beobachtungsdaten

die Parameter der 10 Operationen mit hinreichender Genauigkeit geschätzt und die Itemparameter des RM zufriedenstellend approximiert werden.

In einer Folgeuntersuchung griff NÄHRER (1977, 1980b) u.a. die gerade hier besonders interessierende Frage auf, ob die gefundenen Ergebnisse unabhängig von der untersuchten Personen- und Aufgabenstichprobe gültig sind. In einer Kreuzvalidierungsstudie konstruierte er nach FORMANN's Regeln (nur) 10 neue Items und verglich anschließend die Übereinstimmung zwischen den aus einem neuen Datensatz geschätzten RASCH-Itemparametern (b) mit den aus FORMANN's Operationsparameterschätzungen prognostizierten Aufgabenschwierigkeiten (t). Differenzierte Modellkontrollen (vgl. dazu auch NÄHRER, 1980a) bestätigten die Hypothese, daß die konstruierten Items homogen im Sinne des RM sind. Ihre vorhergesagten Schwierigkeiten t unterscheiden sich allerdings zum Teil erheblich von den Itemparameterschätzung b. Die hier deutlich werdende Stichprobenabhängigkeit des Modells zeigte sich auch beim unmittelbaren Vergleich der Operationsparameterschätzungen aus den beiden Datensätzen. NÄHRER führt die signifikant unterschiedlichen Schwierigkeitsschätzungen für einige der angenommenen Operationen auf Unterschiede in der Strukturfestlegung bei bestimmten Items, d.h. mit anderen Worten, auf mangelnde Objektivität/ Operationalisierung der Konstruktionsregeln zurück. Er modifizierte deshalb einige Konstruktionsregeln, eliminierte die Faktorstufe "Anzahl" und konnte so verbesserte Modellanpassungen in beiden Datensätzen erreichen.

Die relativ geringe Zahl der letztlich analysierten Items (k=12) im Vergleich zur Anzahl geschätzter Operationsparameter (m=9, wobei pro Faktor ein Parameter, also insgesamt 3, auf 0 normiert wurden), verbietet allerdings hier schon von einer hinreichend validierten Konstruktrepräsentation zu sprechen.

7.2 WEITERE UNTERSUCHUNGEN ZUR ENTWICKLUNG EINES KONSTRUKTIONSANSATZES FÜR ALGORITHMISCHE DENKAUFGABEN

In den Untersuchungen von FORMANN (1973) und PISWANGER (1975), der das Itemmaterial einer großen heterogenen Probandenstichprobe vorlegte, wurden 9 bzw. 18 (!) von 42 Matrizenaufgaben wegen mangelnder LLTM-Konformität selegiert. Obwohl die Gründe der Inhomogenität nicht immer eindeutig zu erkennen sind, legten Strukturanalysen der ausgeschiedenen und verbliebenen Items nahe, bei der Konstruktion darauf zu achten, (1) die Figuren graphisch exakt darzustellen und auf solche Figuren gänzlich zu verzichten, die eine intendierte Operation durch zusätzlich erforderliche Wahrnehmungsleistungen für einen Teil der Pbn erschweren, (2) relevante Materialeigenschaften leicht unterscheidbar zu machen ('Vorder- und Hintergrund') und (3) 'Diagonallösungen', die von einigen Pbn bevorzug werden, nicht attraktiv erscheinen zu lassen. Um die Subjektivität der Itemkonstruktionen weiter zu reduzieren, schien es darüber hinaus ratsam, die bei FORMANN (1973) nicht spezifizierte, quasi unendlich große Menge möglicher Repräsentanten für die verschiedenen Materialeigenschaften auf eine überschaubare Menge graphisch leicht realisierbarer Repräsentanten zu reduzieren. Gleichermaßen sollte bei den Distraktoren dafür gesorgt werden, daß "richtige" und "falsche" Komponenten gleichmäßig verteilt werden (s. dazu auch Kp. 7.4).

Nachfolgend sollen die Ergebnisse zweier Versuche rationaler Aufgabenkonstruktionen, die, ausgehend von den revidierten Konstruktionsregeln NÄHRER's (1977, 1980b), auch die o.a. Konstruktionsvorschläge einbeziehen, als Ausgangspunkt für die Weiterentwicklung der Konstruktionstheorie zusammenfassend dargestellt werden.

Um der Frage nachgehen zu können, ob auf der Grundlage des von NÄHRER modifizierten Modells des Lösungsverhaltens von Matrizen-

aufgaben und den dort publizierten Operationsparameterschätzungen, homogene Items mit (zuvor) bestimmten Schwierigkeiten synthetisierbar sind, konstruierte HABON (1981) 14 neue Matrizenaufgaben zu 7 avisierten Schwierigkeitsstufen (s. auch HORNKE und HABON, 1982). Aus einer Liste von 830 Itemstrukturen mit ihren jeweils vorhergesagten Schwierigkeiten t, wurden dabei zu den 7 festgelegten Schwierigkeitsstufen, die Aufbauregeln von je 2 Items ermittelt und der Konstruktion zugrundegelegt.

Die so erstellten 14 Matrizenaufgaben wurden zusammen mit den 12 Items, die NÄHRER's Operationsparameterschätzung zugrundeliegen, 322 Probanden im Alter von 17 bis 37 Jahren ohne Zeitbegrenzung zur Bearbeitung vorgelegt. Die Konstrollen zum RASCH-Modell, dessen Gültigkeit Voraussetzung für die LLTM-Interpretierbarkeit ist, legen den Schluß nahe, daß durch den weiter "reglementierten" Konstruktionsansatz weitgehend homogene Matrizenitems entwickelt werden können. Das LLTM ist in der von NÄHRER übernommenen Form aber nur dann interpretierbar und für die angezielten Zwecke technologisch brauchbar, wenn die Schwierigkeit aller homogenen Aufgaben hinreichend approximiert werden kann. Die graphische Gegenüberstellung vorhergesagter (t(g)) und geschätzter Itemschwierigkeiten (b(g)) vermittelt jedoch ein uneinheitliches Bild (s. Abbildung 5, nächste Seite).

Es fallen hier nicht nur deutliche Fehlprognosen bei den hinzukonstruierten Aufgaben 17 und 22 auf, sondern wider Erwarten auch einige Abweichungen bei den übernommenen Aufgaben, die ja den Operationsparameterschätzungen NÄHRER's zugrundelagen (7,12,6,9). Ein direkter Vergleich der von NÄHRER (1977) berichteten Schwierigkeitsschätzungen b mit den entsprechenden hier ermittelten Werten zeigt, daß beispielsweise die Items 7 und 12 bei NÄHRER eine weitaus geringere Schwierigkeit aufweisen als im vorliegenden Fall. Die mangelnde Invarianz der RASCH-Parameter ist vermutlich auf Stellungseffekte der Items zurückzuführen, die das LLTM in vorliegender Form aber nicht berücksichtigt. Der zusätzliche

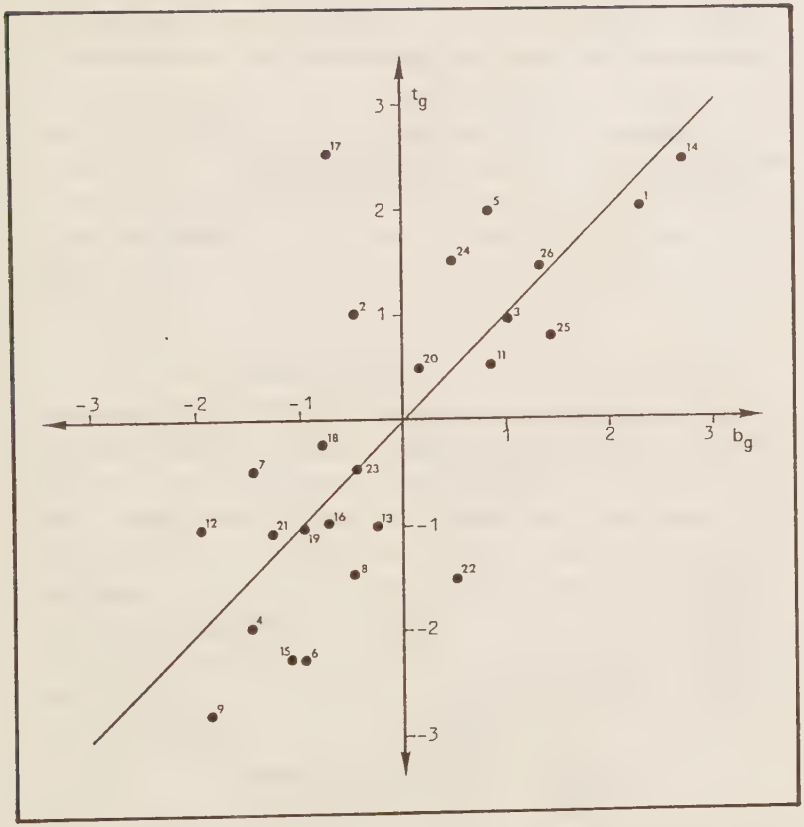


Abbildung 5 :
Graphische Modellkontrolle zur Überprüfung der Übereinstimmung geschätzter Schwierigkeits-parameter b_g und prognostizierter Itemschwierigkeiten t_g (vgl. HABON, 1981)

Vergleich der Operationsparameterschätzungen aus beiden Datensätzen (vgl. Abb 6; die Operationen 'Form', 'überlagern' und 'waagerecht' wurden auf Null normiert), zeigt, daß die Schätzungen für die Operationen 's=senkrecht' und 'F=Fortsetzen' hier nicht repliziert werden.

Abb.

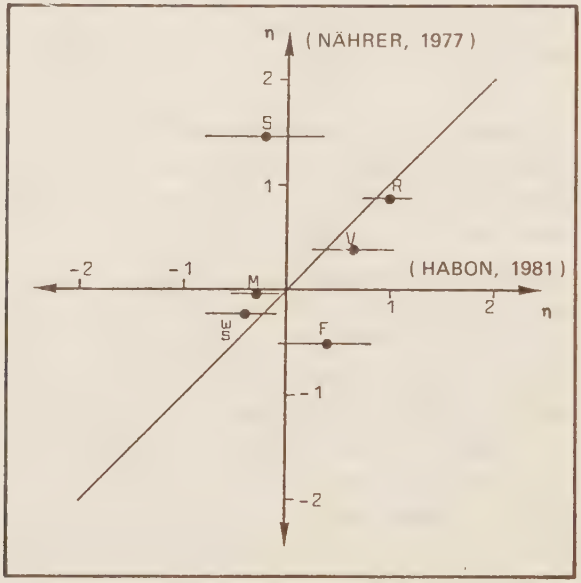


Abbildung 6 :
Vergleich der Operationsparameterschätzungen
bei gleicher Itemmenge und verschiedenen
Personenstichproben (vgl. NÄHRER, 1977; HABON, 1981)

Diese Abweichungen sind nun aber leicht durch die beobachteten Inhomogenitäten erklärbar, die sich unter den hier vorliegenden ungünstigen Schätzvoraussetzungen für die Operationsparameter (UNTERrepräsentationen einiger Operationen in den Aufgaben), stark verfälschend auf die Schätzungen der Operationsschwierigkeiten auswirken. Abhängigkeiten von der Zusammensetzung der Aufgabenstichproben führen nun bei den Schwierigkeitsprognosen hinzukonstruierter Items zu durchgängigen Unterschätzungen der Schwierigkeit jener Aufgaben, die aufgrund der Modellannahmen

eher als relativ 'leicht' einzuschätzen sind (8,15,22, s. Abb.5) und zur teilweisen Überschätzung 'schwieriger' Aufgaben (5,17,24). Dieser Effekt schlägt sich auch in hochsignifikanten Prüfgrößen von Likelihoodquotiententests nieder, die zur Ablehnung des Modells in der Version NÄHRER's führen. Nach inhaltlich begründeter Selektion einiger Items konnten unter Zugrundelegung aller übrigen Aufgaben (k=22) interpretierbare und hinreichend präzise Operationsparameterschätzungen gewonnen werden. Es bedarf allerdings wiederum einer erneuten Kreuzvalidierung mit neuen Pbn und Items, um die theoretisch und praktische Relevanz der Schätzungen nachzuweisen.

In einer Folgeuntersuchung stellten HORNKE und HABON (1983), entsprechend der Vermutung, daß Items mit gleicher Operationsstruktur, aber verschiedenem Erscheinungsbild auch vergleichbare Schwierigkeiten aufweisen, vier regelgeleitet konstruierte Paralleltests zusammen. Jeder Test bestand aus 11 Matrizenaufgaben und zusätzlich drei Ankeritems, die in allen vier Versionen vorkommen. Einer der vier Tests setzte sich aus Items zusammen, die schon von HORNKE und HABON (1982) hinsichtlich ihrer RASCH-Homogenität untersucht worden waren. Alle Aufgaben wurden nach den von HABON (1981) spezifizierten Regeln konstruiert und ihre Schwierigkeit mittels der dort gefundenen Operationsparameterschätzungen vorhergesagt.

Die insgesamt 563 Probanden vorgelegten Aufgaben unterschieden sich allerdings von den bisher analysierten Items in einem Punkt: es waren keine Antwortalternativen vorgegeben, sondern der Proband hatte die Lösung selber einzuzeichnen. Hierdurch sollte der wahrscheinliche Einfluß der Wahlantworten auf den Lösungsprozeß ausgeschaltet und ein aktives Antwortkonstruieren erzwungen werden.

Kontrollen zur Lösung des RM wiesen einige Items als modellunverträglich aus. Diese Inhomogenitäten ließen sich allerdings auf

Auswerteunsicherheiten bzw. drucktechnische Mängel zurückführen. Korrelationen über .80 zwischen vorhergesagten und geschätzten Itemschwierigkeiten bei allen vier Versionen zeigen, daß bei homogenen Aufgaben die Itemparameter befriedigend aus den Operationsparametern vorhergesagt werden konnten. Die Itemparameter der früher mit Distraktoren vorgelegten Aufgaben erwiesen sich als beachtlich stabil (r=.91). Aus den Daten wurden erneut die Schwierigkeitsparameter der neun Operationen geschätzt. Die Operation 'Variieren' scheint in diesen Aufgaben schwerer zu sein als in den zuvor untersuchten Items, die Operation 'waagerecht/senkrecht' dagegen leichter.

Ein Überblick über die Operationsparameterschätzungen der verschiedenen, unabhängigen Studien zum vorliegenden Konstruktions-/und Denkmodell zeigt die relativ hohe Stabilität und damit itemstichprobenunabhängige Gültigkeit der Schätzungen für die meisten der angenommenen Operationen. Es ist zu beachten, daß vor 1977 eine erweiterte Modellversion galt. Die Schwierigkeitsbeiträge der Operationen 'Variieren' und 'waagerecht/senkrecht' scheinen dagegen stärker von den jeweiligen Zusammensetzungen der Itemstichproben abhängig zu sein.

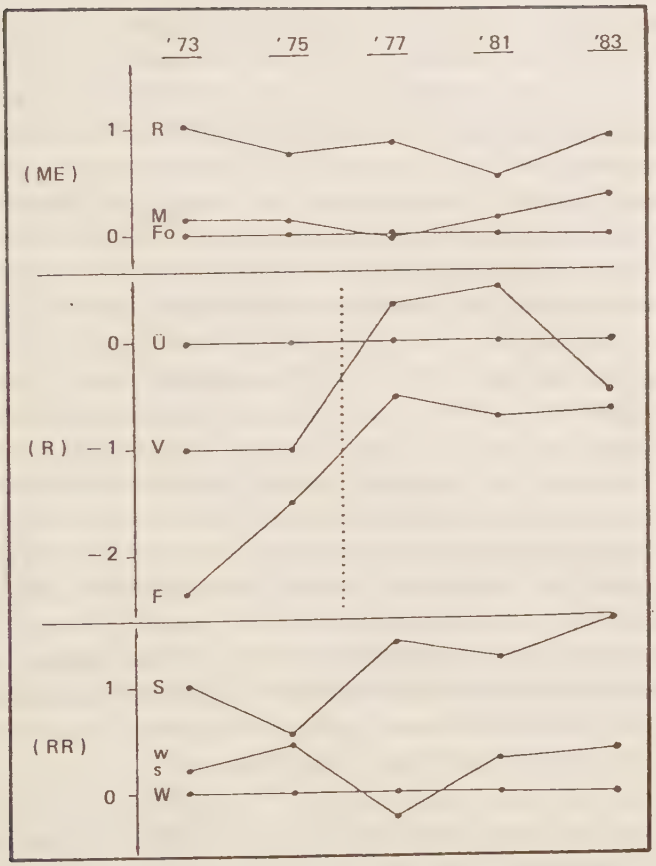


Abbildung 7 :
Vergleich der Operationsparameterschätzungen des ursprünglichen und modifizierten LLTM aus den Untersuchungsdaten von HORNKE & HABON (1983) und den Analysen FORMANNs (1973), PISWANGERs (1975), NÄHRERs (1977) und HABONs (1981). " ME " bezeichnet Materialeigenschaften, " R " Regeln und " RR " Richtungen der Regelgeltung.

7.3 ERWEITERUNG UND REVISION DES KONSTRUKTIONSMODELLS

Die Ergebnisse der nunmehr fünf unabhängigen Analysen zu der von FORMANN (1973) aufgestellten Hypothese über die Komponenten des Lösungsverhaltens bei Matrizenitems lassen im Rückblick zweifellos eine deutliche Verbesserung der Modellkonformität und eine gewisse Stabilisierung der Operationsparameterschätzungen mit zunehmender Ein- und Abgrenzung der Konstruktionsprinzipien erkennen. Dennoch läßt eine ganze Reihe aufscheinender Probleme die Frage angemessen erscheinen, ob der vorliegende Konstruktionsansatz die oben gestellten Anforderungen an die hier dringend erforderliche Itemkonstruktionstechnologie erfüllt.

Ein wesentliches und bisher nicht gelöstes Problem besteht in der nachgewiesenen Itemstichprobenabhängigkeit der Schwierigkeitsschätzungen einiger Operationen. Sie kann beispielsweise dazu führen, daß bei der rationalen Konstruktion von Aufgaben, deren Schwierigkeiten nicht in dem Skalenbereich liegen soll, aus dem die Aufgaben stammen, die zur Operationsparameterschätzung herangezogen wurden, systematisch "leichte" Items in ihrer Schwierigkeit vom Denkmodell UNTERschätzt, besonders aber die Komplexität "schwerer" Aufgaben ÜBERschätzt wird (vgl. HABON, 1981). Die Abhängigkeit der Operationsparameterschätzungen vom Komplexitäts-GRAD der Aufgabenstichprobe läßt die theoretisch implizierte Spreizbarkeit der Skala fraglich erscheinen.

Die bisherigen Ergebnisse deuten darauf hin, daß mit dem zuletzt doch recht stark eingegrenzten Regel-/ bzw. Operationsinventar Aufgaben von der Komplexität, wie sie für viele überdurchschnittliche Problemlöser angemessen erscheint, ohne weiteres NICHT konstruierbar sind. Die Gründe hierfür dürften einerseits in einer zunehmenden Exhaustion und Spezifikation des Operationssatzes im Verlauf der verschiedenen Validierungen zu suchen sein, andererseits aber auch darin, daß die Operationen eben Lösungsprozesse mittelschwerer bis leichter "Probleme" konstituieren.

Die Einschränkung des Konstruktionsregelsatzes erhöht scheinbar die Validität der Homogenitäts- und Schwierigkeitsprognosen für Aufgaben eines EINGEGRENZTEN Schwierigkeitsbereichs, sie bringt aber andererseits mehrere Nachteile mit sich:

(1) Bei einer insgesamt wenig differenzierten Operationsstruktur sind auch die Aufgaben in ihrem Anforderungsgehalt einander sehr ähnlich. Im Lösungsprozeß der Aufgaben werden einzelne Operationen mitunter häufig wiederholt UND geübt, d.h. daß ihre Schwierigkeit im Verlauf der Testbearbeitung abnehmen kann. In diesem Zusammenhang weisen FISCHER und FORMANN (1982) auf die Problematik von Lerneffekten für das adaptive Testen hin, da hier ja homogene Aufgaben vorausgesetzt werden, deren Schwierigkeiten konstant sind und nicht von der Übung ihrer Lösungsoperationen in vorangehenden Aufgaben abhängen. Die in den zitierten Untersuchungen beobachteten Stellungseffekte lassen einen "operationsspezifischen Übungstransfer" (vgl. SCHEIBLECHNER, 1972; SPADA, 1976) bei den bislang konstruierten Matrizenaufgaben vermuten. Ungeklärt sind bisher auch die möglichen Auswirkungen relativ "eintöniger" Problemstrukturen auf die Testmotivation. Hier könnte man demotivierende Wirkungen vermuten, die den möglichen Übungseffekten nach Bearbeitung einer größeren Zahl von Items entgegenwirken oder sie überlagern.

(2) Die mit dem Wegfall bzw. der Spezifikation einiger Konstruktionsregeln/ Operationen verbundene inhaltlich definitorische Einschränkung des Latent Trait wirkt sich zudem vermutlich negativ auf seine diagnostische Nützlichkeit aus. Es ist anzunehmen, daß ein "Intelligenztest" mit hochdifferenzierter und breitgefächerter kognitiver Anforderungsstruktur in mehr Situationen bessere Erfolgs- bzw. Mißerfolgsprognosen und -erklärungen erlaubt, als ein Test, der lediglich eine, was Art und Zahl erforderlicher Lösungsoperationen betrifft, eingeschränkte Fähigkeit mißt.

(3) Bezüglich der oben formulierten Konstruktionsziele ist jedoch entscheidend, daß mit dem vorliegenden Regelsatz Items "hoher" Schwierigkeit im Sinne "klassischer" Lösungswahrscheinlichkeiten nicht konstruiert werden können. Die bei Gültigkeit des Modells naheliegende Möglichkeit, die Zahl der Operationen bzw. Operationswiederholungen in einer Aufgabe zu erhöhen, um sie in vorhersagbarer Weise schwieriger zu machen, ist dann nicht zielführend, wenn die Schwierigkeitsschätzungen einzelner Operationen in unbekannter Weise von der Gesamtkomplexität der jeweils analysierten Items abhängen. Dies scheint aber für die Matrizenitems zuzutreffen, wie teilweise deutliche Überschätzungen der tatsächlichen Itemschwierigkeiten durch das LLTM zeigen.

Der vorliegende Ansatz hat sich für die Konstruktion weitgehend homogener Aufgaben eines begrenzten Anforderungsgehalts als brauchbar erwiesen. Die o.a. Schwächen des Konstruktionsmodells machen aber gerade vor dem Hintergrund der hier formulierten Zielvorstellungen eine Revision und Erweiterung des Ansatzes erforderlich.

Ein Ziel der Erweiterung des Konstruktionsmodells soll nunmehr darin bestehen, mehr Regeln/ Operationen zu finden, die Matrizenaufgaben (noch) schwieriger machen und dabei zugleich mit hoher Wahrscheinlichkeit Elemente einer eindimensionalen Problemlösungsfähigkeit sind. Diese neuen Operationen sollen sich zwanglos in das bestehende Konstruktionssystem einordnen lassen und zusammen mit den "bewährten" Operationen eine latente Dimension inhaltlich definieren.

Die Suche nach "neuen" Operationen läßt mehrere Vorgehensweisen offen. Ein möglicher und bislang wohl häufig beschrittener Weg besteht darin, Regeln/ Operationen zu ERFINDEN, von denen man mutmaßt, das sie irgendetwas mit dem fraglichen Merkmal zu tun haben und die Lösung entsprechend konstruierter Items erschweren.

Ein solches Vorgehen ist allerdings bei der Frage nach der Dimensionalität der letztlich konstruierten Items auf bloße Vermutungen angewiesen. Vielversprechender scheint dagegen zu sein, in vorhandenen (eindimensionalen) Intelligenztests nach ergänzenden Konstruktionsregeln bzw. Operationen zu suchen. Da es um solche Operationen gehen soll, die einen eigenständigen und größeren Schwierigkeitsbeitrag liefern, wurden hier die Aufgaben des Set II der ADVANCED Progressive Matrizes (APM) von RAVEN (1962) ausgewählt und hinsichtlich ihres strukturellen Aufbaus analysiert.

Ein Teil der 36 APM-Items ist mit den bekannten Aufbauregeln vollständig beschreibbar, ein anderer hingegen nicht. Eine hypothetische Strukturmatrix auf der folgenden Seite gibt an, wie häufig welche der bekannten 'Materialeigenschaften', 'Regeln' und 'Richtungen' der Regelgeltung in den Aufgaben der APM erkannt werden müssen, um sie zu lösen. Erwarungsgemäß können die nach steigender Schwierigkeit angeordneten Items der ERSTEN Testhälfte beschrieben werden. Die "schweren" Items der zweiten Testhälfte enthalten hingegen Regeln/ Operationen, die durch die vorliegende Konstruktrepräsentation nicht abgebildet werden.

Um diese hier besonders interessierenden Operationen identifizieren, klassifizieren und hinsichtlich ihrer relativen Bedeutsamkeit (=Auftretenshäufigkeit) besser einschätzen zu können, wird im folgenden der Versuch einer Strukturbeschreibung der APM mit einigen aus der Literatur bekannten, konkurrierenden Klassifikationsansätzen unternommen.

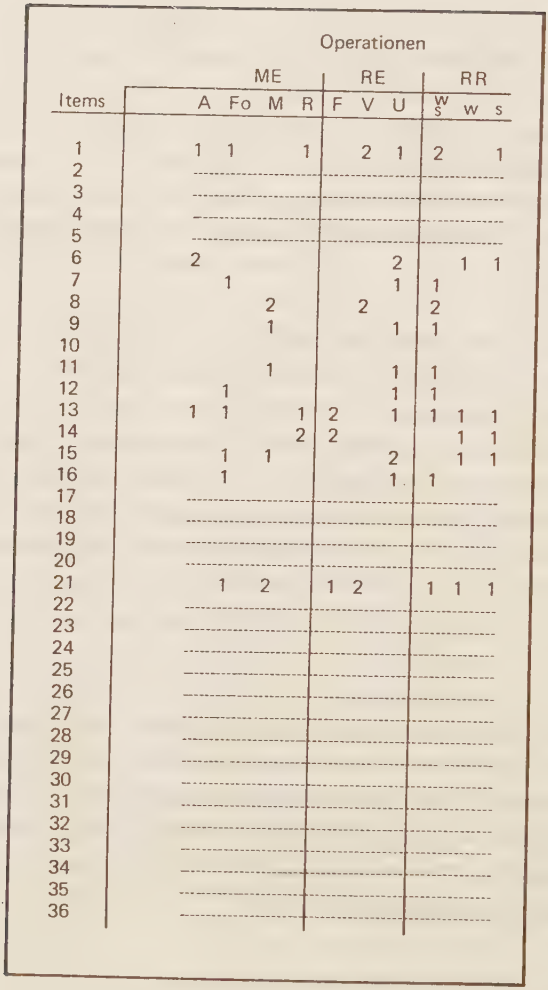


Abbildung 8 :
Versuch einer Strukturbeschreibung der 36 APM – Items mit dem Operationsinventar von FORMANN (1973)

7.3.1 DAS BESCHREIBUNGSSYSTEM VON JACOBS UND VANDEVENTER (1972)

Auf der Suche nach repräsentativen Aufgaben, die sich einerseits für ein Training der durch die Kategorienparameter 'CFR' definierten Teilfähigkeit: 'Erkennen figuraler Relationen' des GUILFORD'schen SI-Modells und andererseits zur Beobachtung von Übungstransfer innerhalb der CFR-Kategorie eignen sollten, stellten die Autoren mangelnde inhaltliche Übereinstimmung zwischen den Matrizenaufgaben fest, die im allgemeinen zu dieser Modellkategorie gezählt werden. Da die Kenntnis der Aufbaugesetzmäßigkeit jener Aufgaben von den Autoren als notwendige Voraussetzung für die Planung und Evaluation eines Trainingsprogramms betrachtet wird, stellt sich für sie die Frage nach dem "Content of Intelligence" (ebd., S.241).

Um für die CFR-Kategorie eine befriedigende Antwort zu finden, analysierten JACOBS und VANDEVENTER die Struktur von 166, im Sixth Mental Measurement Yearbook (BUROS, 1965) aufgeführten Gruppenintelligenztests sowie 35 weitere Tests aus der Sammlung des Educational Testing Service und stießen bei 22 Tests auf insgesamt 1.335 CFR-Items. Eine Strukturanalyse ergab, daß bei dem größten Teil dieser Items über die Itemelemente lediglich 12 unterscheidbare Relationen definiert sind (s. nächste Seite).

Der FORMANN'sche Regelkatalog geht nahezu vollständig in diesem Beschreibungssystem auf: Den Operationen 'shape', 'shading', 'movement in a plain', 'flip-over' und 'number series' entsprechen die Materialeigenschaften 'Form', 'Muster', 'räumliche Anordnung' und 'Anzahl' in Kombination mit einer der 'geltenden Regeln'. 'Identity', 'addition' und 'elements of a set' entsprechen den Regeln 'Fortsetzen', 'Überlagern' und 'Variieren'. Weitet man die Definition von 'elements of a set' dahin aus, daß einzelne Figuren nicht nur in Bezug auf die Position im Schema, sondern auch hinsichtlich anderer Dimensionen, wie Größe, Anzahl etc. variieren, ohne aber dabei die Elementenklasse (Dreiecke,

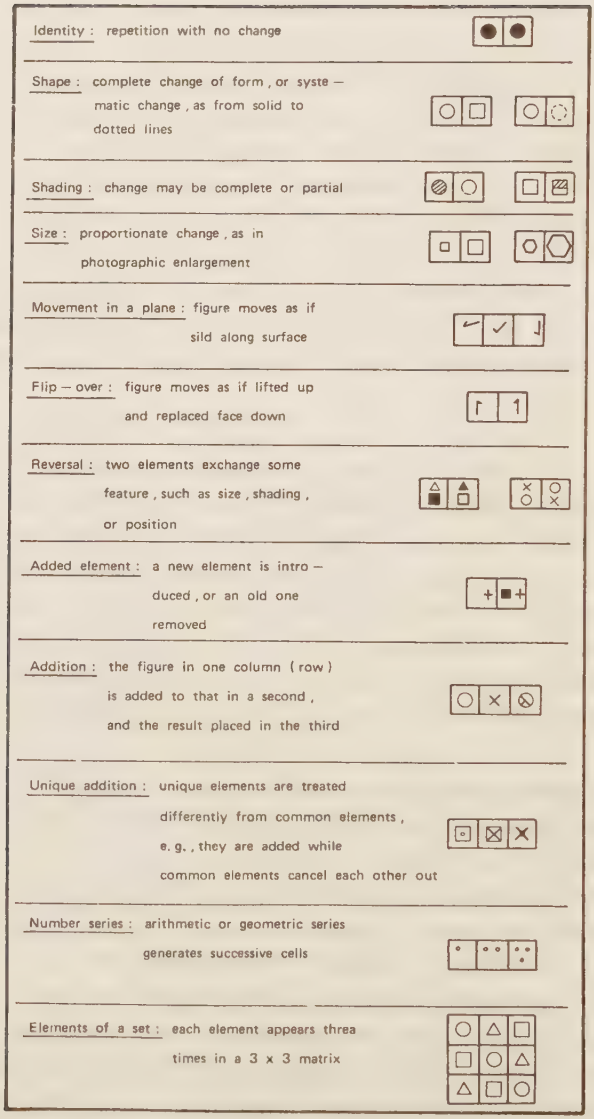


Abbildung 9 :
Liste der 12, in CFR – Items vorfindbaren Relationen;
entnommen aus JACOBS und VANDEVENTER (1972, 243)

Kreise, Linien) zu verlassen, sind 28 der 36 APM-Items klassifizierbar.

PUTZ-OSTERLOH (1981) kritisiert das Beschreibungssystem, indem sie darauf hinweist, daß die Klassifikationsmerkmale nicht disjunkt sind, so z.B. 'shading', das durch 'addition' oder 'added elements' verändert werden kann. Außerdem seien "das eine mal die Merkmale und das andere mal die Veränderungen Einteilungsgesichtspunkte für die unterschiedlichen Relationen" (S. 59). Ein wesentlicher Nachteil des Ansatzes ist sicher auch darin zu sehen, daß die Richtung der Regelgeltung unbeachtet bleibt.

7.3.2 DAS KLASSIFIKATIONSSYSTEM VON WARD UND FITZPATRICK (1973)

Auch WARD und FITZPATRICK (1973) analysierten die Charakteristika von Matrizenaufgaben. Ihr Interesse an einer Beschreibungsmöglichkeit erwuchs aus einer Untersuchung mit Matrizenaufgaben, bei denen die Antworten von den Probanden selbst einzuzeichnen waren.

> "It was than found, for instance, that item difficulty tended to rise oder fall as a function of the number of all attributes within the matrix and the complexity of operations required for solution. The intriguing possibility was raised of being able to distinguish effectively between items in terms of their content and process" (S. 988).

WARD und FITZPATRICK unterscheiden isoklinische (= symmetrische) und synklinische (= asymmetrische) Matrizenaufgaben. Figur 1 in Abbildung 10 (s. nächste Seite) zeigt isoklinische Items, wie sie im Set I der APM vorkommen, Figur II und Figur III verdeutlichen die schon bekannten Operationen 'Fortführen' (=identity) und 'Variieren' (=latin square) an einigen Beispielen.

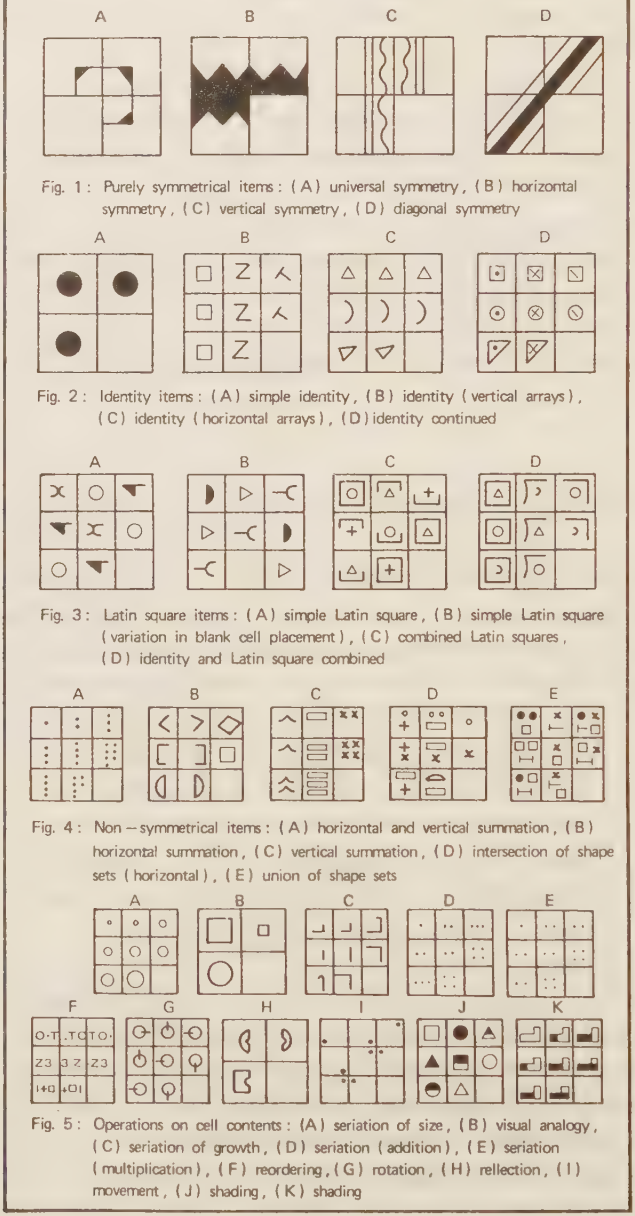


Fig. 1: Purely symmetrical items: (A) universal symmetry, (B) horizontal symmetry, (C) vertical symmetry, (D) diagonal symmetry

Fig. 2: Identity items: (A) simple identity, (B) identity (vertical arrays), (C) identity (horizontal arrays), (D) identity continued

Fig. 3: Latin square items: (A) simple Latin square, (B) simple Latin square (variation in blank cell placement), (C) combined Latin squares, (D) identity and Latin square combined

Fig. 4: Non-symmetrical items: (A) horizontal and vertical summation, (B) horizontal summation, (C) vertical summation, (D) intersection of shape sets (horizontal), (E) union of shape sets

Fig. 5: Operations on cell contents: (A) seriation of size, (B) visual analogy, (C) seriation of growth, (D) seriation (addition), (E) seriation (multiplication), (F) reordering, (G) rotation, (H) reflection, (I) movement, (J) shading, (K) shading

Abbildung 10:
Beispiele für die Klassifikation von
WARD und FITZPATRICK (1973)

Asymmetrische Items enthalten die bekannte Operation 'Überlagern' (=Summation), darüber hinaus auch 'Disjunktionen' (=intersection of shape sets), 'Konjunktionen' (=Union of shape sets) und verschiedene Formen von Seriationen (=seriation of size, growth, etc.). Die Autoren vermuten, daß die nach diesem Schema identisch geschriebenen Items die gleiche Schwierigkeit aufweisen, überprüft wurde diese Annahme allerdings nicht.

7.3.3 DIE BESCHREIBUNG VON PUTZ-OSTERLOH (1981)

PUTZ-OSTERLOH (1981) hält die vorliegenden Klassifikationssysteme zur Beschreibung der Items aus dem Set II der APM für unzureichend, da sie alle nur einzelne Relationen berücksichtigen, in vielen Items aber gleichzeitig mehrere solcher Relationen auftreten und berücksichtigt werden müssen. In ihrem eigenen Beschreibungsschema sollen weder "die Art der aufzudeckenden Relationen und auch nicht die Anzahl anzuwendender Operationen zur Analyse der Relationen" (S. 86) Unterscheidungskriterien für die Items bilden, sondern "die Art des Analyseverfahrens zum Finden der Relationen" (S. 87).

PUTZ-OSTERLOH definiert die vier Kategorien: (1) Dimensionen und Valenzen, (2) kontinuierliche Veränderung, (3) abhängige Veränderung und (4) merkmalsabhängige Veränderung.

Die beiden Schemata der Items 1 und 29 der APM in Abbildung 11 verdeutlichen, was unter "Dimensionen und Valenzen" zu verstehen ist. In Item 1 verändern sich die Figuren in den Dimensionen 'Form', 'Anzahl' und 'räumliche Lage'. Die Dimension 'Form' hat die Valenzen 'Kreis', 'stehendes' und 'liegendes Quadrat', die Dimension 'Anzahl' die Valenzen '1', '2' und '3' und die Dimension 'räumliche Lage' die Valenzen 'rechts geneigt', 'links geneigt' und 'senkrecht'. Auch Item 29 besteht aus drei Dimensionen

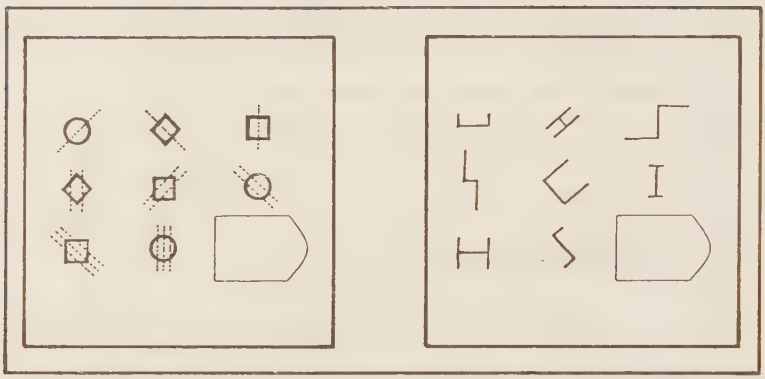


Abbildung 11 :
Beispielitems mit verschiedenen " Dimensionen und Valenzen ";
links : Item 1 , rechts : Item 29 der APM von RAVEN (1962)

mit je 3 Valenzen. Die unterschiedliche Stellung in der nach aufsteigenden Schwierigkeiten geordneten Itemreihe deutet darauf hin, daß Aufgabe 29 wesentlich seltener richtig gelöst wird als Aufgabe 1. Allein an diesem Beispiel wird deutlich, daß die in beiden Items zu erkennenden Dimensionen und Valenzen (je 3) offenbar keine ausreichend differenzierte Strukturbeschreibung darstellen, um die unterschiedlichen Itemschwierigkeiten zu erklären.

Die Kategorie 'kontinuierliche Veränderung' entspricht etwa der 'seriation' bei WARD und FITZPATRICK sowie dem 'flip-over' und 'movement in a plain' bei JACOBS und VANDEVENTER. Was PUTZ-OSTERLOH als 'abhängige Veränderung' bezeichnet, findet seine Entsprechung in der Operation 'addition' und 'unique addition' bei JACOBS und VANDEVENTER, 'summation' bei WARD und FITZPATRICK

und 'Überlagern' bei FORMANN. Mit der Kategorie 'merkmalsabhängige Veränderungen' gelingt es PUTZ-OSTERLOH nun aber auch die Aufgaben zu klassifizieren, die nach zuvor referierten Ansätzen nicht beschrieben werden können. Es handelt sich hierbei um Items, bei denen für verschiedene Merkmale (=Repräsentanten) der Figuren unterschiedliche Regeln gefunden werden müssen. Die Bindung der Regeln an BELIEBIGE graphische Repräsentanten hat zur Folge, daß keine MaterialUNabhängigen (durchgängigen) Konstruktionsregeln/ Operationen definierbar sind. Items dieser Art können daher für vorliegende Zwecke nicht berücksichtigt werden.

7.3.4 DIE ERWEITERUNG DES KONSTRUKTIONSANSATZES VON FORMANN

Vergleicht man die referierten Strukturbeschreibungen miteinander, so fällt auf, daß sie sich in weiten Teilen überschneiden und neben den Konstruktionsregeln FORMANN's, die sich in nahezu allen Ansätzen finden lassen, zusätzliche Regeln bzw. Operationen enthalten, von denen zu vermuten ist, daß sie schwierigkeitserzeugend wirken.

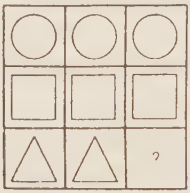
Das folgende Regel-/ bzw. Operationsinventar stellt einen Versuch dar, diese Regeln in das vorhandene Beschreibungs- und Konstruktionssystem zu integrieren.

Es sind folgende Regeln (RE) zu unterscheiden, die die Beziehung zwischen den 8 Figuren des Schemas festlegen:

Je Zeile und/oder Spalte des Schemas

--- ist dreimal derselbe Repräsentant angeordnet.

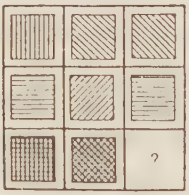
- IDENTITÄT (I)



Die Operation IDENTITÄT entspricht den Operationen 'Fortsetzen' (FORMANN) und 'identity' (JACOBS und VANDEVENTER, WARD und FITZPATRICK).

--- entsteht ein Repräsentant einer Komponente dadurch, daß die beiden restlichen, in derselben Zeile/oder Spalte angeordneten Repräsentanten, übereinander gelegt gedacht werden.

- ADDITION (A)

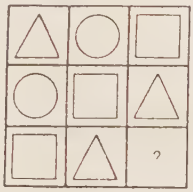


(='addition': JACOBS und VANDEVENTER/ 'summation': WARD und FITZPATRICK / 'Überlagern': FORMANN)

Je Zeile und/oder Spalte des Schemas

--- variiert die Reihenfolge von drei Repräsentanten.

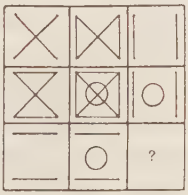
- VARIIEREN (V)



(='latin square': WARD und FITZPATRICK / 'elements of a set': JACOBS und VANDEVENTER / 'Variieren': FORMANN)

--- entsteht ein Repräsentant einer Komponente dadurch, daß von einer vollständigen Figur (in der ersten oder zweiten Zeilen-/ oder Spaltenposition eine TEILfigur ausgeblendet wird.

- SUBTRAKTION (SU)



(='removed elementes': JACOBS und VANDEVENTER)

--- entsteht ein Repräsentant einer Komponente dadurch, daß von den beiden restlichen, in derselben Zeile oder Spalte angeordneten Repräsentanten nur die Elemente übernommen werden, die sowohl in der einen als auch in der anderen Figur vorkommen.

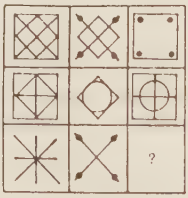
- DISJUNKTION (D)

(='intersection': WARD und FITZPATRICK)

Je Zeile und/oder Spalte des Schemas

--- entsteht ein Repräsentant einer Komponente dadurch, daß von den beiden restlichen, in derselben Zeile oder Spalte angeordneten Repräsentanten nur die Elemente übernommen werden, die jeweils nur in der einen, nicht aber in der anderen Figur vorkommen, wobei es jedoch in beiden Figuren auch gemeinsame Elemente gibt.

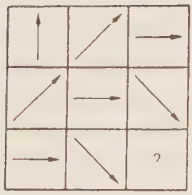
- EINZELELEMENTENADDITION (EA)



(='unique addition': JACOBS und VANDEVENTER)

--- entsteht ein Repräsentant einer Komponente dadurch, daß die Veränderungsregel zur Umwandlung der ersten in die zweite Figur (Verschiebung in der Ebene oder im Raum; Addition oder Subtraktion) auf die veränderte zweite Figur angewendet wird um die dritte Figur zu erzeugen.

- SERIATION (SE)



(='seriation': WARD und FITZPATRICK / 'flip-over', 'movement in a plain': JACOBS und VANDEVENTER)

Die auf der Grundlage dieser Regeln konstruierten Denkaufgaben können als graphische Repräsentanten des logischen Aussagenkalküls aufgefaßt werden (vgl. dazu auch SCHEIBLECHNER, 1972). Erweisen sie sich als homogen im Sinne eines formalen Testmodells, definieren die Operationen einen Latent Trait, der als "AUSSAGENLOGISCHE DENKFÄHIGKEIT" bezeichnet werden kann. Damit wäre ein aufgaben- bzw. materialunabhängiges Konzept gegeben, das die rationale Konstruktion von Matrizenitems substanztheoretisch begründet.

SCHEIBLECHNER (1972), der eben diese Fähigkeit mit phänomenologisch anderen, sogenannten "Mengenrechenaufgaben" erfaßt, weist, wie MISPELKAMP (1985), auf die Bedeutung der übergeordneten theoretischen Orientierung hin:
"Gelingt hierfür eine befriedigende psychologische Theorie, so hat dies den entscheidenden Vorteil, daß durch (bloß) logische Analyse jeden Denkproblems klargestellt werden kann, welche

Grundoperationen und welche Kombinationen von Operationen für den Lösungsweg notwendig sind. Die Kombinationsregeln, der Definitionsbereich der Theorie und ihre Anwendung bei der Vorhersage des Lösungsverhaltens bei neuen Problemen sind eindeutig festgelegt" (S. 478).

7.3.5 DIE REVISION DES REGELSATZES VON FORMANN UND NÄHRER

Die zitierten Strukturbeschreibungen wie auch die empirischen Befunde zur Validität, d.h. hier zu Stabilität von Schwierigkeitsbeiträgen der Einzeloperationen, legen eine erneute Bewertung des FORMANN'schen Ansatzes nahe. Um die innere Konsistenz des Konstruktionsregelsatzes zu erhöhen und die Operationen zwanglos als Denkoperation interpretieren zu können, soll im folgenden auf die Materialeigenschaften als eigenständiger Faktor im Regelsatz verzichtet werden. Folgende Gründe sprechen dafür:

(1) Die Materialeigenschaften 'Muster' und 'Form' sind nur schwer als disjunkte Klassen definierbar. Ein wenig differenziertes 'Muster' kann ebensogut als 'Form', wie eine komplex strukturierte Form als 'Muster' aufgefaßt werden.

(2) Beide Materialeigenschaften leisten nur einen GERINGEN und relativ INSTABILEN Schwierigkeitsbeitrag.

(3) Die Materialeigenschaft 'räumliche Anordnung' unterscheidet sich von 'Muster' bzw. 'Form' oft nur durch die Wahl bestimmter graphischer Repräsentanten (Striche, Pfeile etc.). Wenn es sich hierbei aber um eine MATERIALSPEZIFISCHE Operation handelt, sollte sie auch entsprechend gekennzeichnet werden.

(4) Was hier noch als Materialeigenschaft 'räumliche Anordnung' verstanden wird, läßt sich problemlos als 'Regel' auffassen. Diskontinuierliche, wie auch kontinuierliche Veränderungen der räumlichen Anordnung einer Figur sind durch die Regeln 'Variieren' bzw. 'Seriation' beschreibbar.

Unter den von FORMANN (1973) konstruierten Aufgaben befinden sich einige, deren Schwierigkeit durch das LLTM durchgängig unterschätzt wird. Es handelt sich dabei um Items, bei denen die Erkennbarkeit relevanter Figurteile als Träger der jeweils

geltenden Regel erschwert ist. PUTZ-OSTERLOH (1981), die sich in ihren denkpsychologischen Untersuchungen zu weiteren Schwierigkeitskomponenten in den APM-Items, gerade auch mit diesem Problem auseinandersetzt, nennt drei Faktoren, die die Erkennbarkeit relevanter Merkmale herabsetzen:

(1) Sie können dann schwer erkennbar sein, wenn sie in einer "geschlossenen Gestalt" enthalten sind. Bei Aufgabe 36 der APM beispielsweise

> "müssen an Figuren aus durchgezogenen Linien die inneren von den äußeren Linien getrennt werden. Hier müssen also Suchoperationen eingesetzt werden, um die variationsrelevanten Teile einer Figur voneinander zu unterscheiden" (S.166).

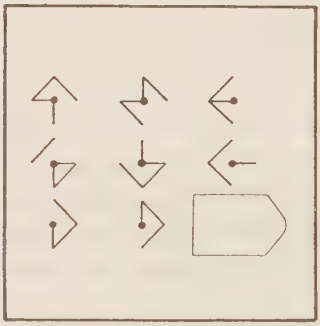


Abbildung 12 :
Erschwerte Erkennbarkeit relevanter Merkmale
im Item 36 der APM von RAVEN (1962)

(2) Bei anderen Aufgaben (vgl. APM Nr. 29 in Abb. 11, S.126), verändert sich die Dimension 'Art der Figur' zeilenweise so stark, daß artgleiche Figuren nur schwer als solche erkennbar sind. PUTZ-OSTERLOH (1981) fand in Blickbewegungsversuchen

heraus,

> "daß bei diesen Aufgaben das Analyseschema gewechselt werden muß, indem nicht nur Zeilen oder Spalten, sondern einzelne dimensionsgleiche Figuren miteinander verglichen werden müssen" (S. 166).

(3) Weiterhin kann die Erkennbarkeit relevanter Merkmale dadurch erschwert sein, daß sie gewisse Charakteristika mit anderen relevanten Merkmalen teilen (Vorder-/ Hintergrund etc.). Bei Aufgaben dieser Art

> "müssen Prozeduren eingesetzt werden, mit deren Hilfe das jeweils analysierte Merkmal hervorgehoben und das störende unterdrückt oder weniger beachtet wird" (S. 167).

Die Ergebnisse dieser Beobachtungen geben wichtige Hinweise darauf, wie MEHRERE Komponenten in EINER Figur realisiert werden und wie sich dies auf das Problemlösungsverhalten der Pbn auswirkt. Die Schlußfolgerung, daß die APM-Items nicht homogen seien, da "bei verschiedenen Items in unterschiedlichem Umfang differierende Problemlösungsschritte den Erfolg der Problembearbeitung mitbestimmen" (PUTZ-OSTERLOH, 1981, S.170), scheint allerdings problematisch, da die Frage der Homogenität eines Itemset die Beantwortung der Frage nach der EMPIRISCHEN Gültigkeit eines Meßmodells voraussetzt, das den Homogenitätsbegriff formalisiert.

Es ist vorstellbar, daß auch zusätzlich erforderliche Suchprozeduren Bestandteile der durch die o.a. Operationen definierten eindimensionalen Fähigkeit sind. Es scheint daher sinnvoll zu sein, unterschiedliche "REALISIERUNGSMODI (RM)" bzw. "SUCHOPERATIONEN" in den Regel-/ Operationssatz zu integrieren:

MEHRERE Komponenten können in EINER Figur auf unterschiedliche Art und Weise realisiert werden:

1. SEPARIERTE KOMPONENTEN (SC) sind durch relevante und KONSTANTE, irrelevante Repräsentantenmerkmale vollständig beschrieben. Außerdem existieren Merkmale (groß versus klein oder Vordergrund versus Hintergrund), die die Unterscheidbarkeit der Komponenten fördern.

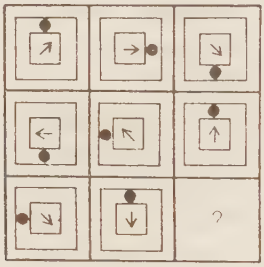


Abbildung 13:
Separierte Darstellung von zwei, jeweils waagrecht geltenden Seriationen (SC-2*"Se"-2*"w")

2. Bei INTEGRIERTEN KOMPONENTEN (IC) ist das relevante Repräsentantenmerkmal einer Komponente ein nicht-konstantes irrelevantes Repräsentantenmerkmal einer anderen Komponente. Hier ist eine Suchoperation erforderlich, wie sie unter Punkt 3 (s.o.) beschrieben wird.

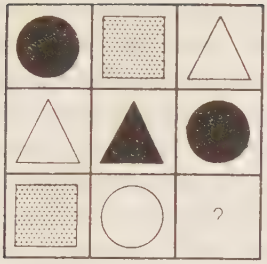


Abbildung 14:
Integrierte Realisationen von zwei, jeweils senkrecht geltenden Variationen

3. INTEGRIERTE REPRÄSENTANTEN (IR) liegen vor, wenn die Komponenten einer Figur durch Repräsentanten realisiert werden, die nur durch die für sie jeweils geltenden Regeln (RE) diskriminierbar sind. Die Repräsentanten bilden eine "geschlossene Gestalt", die von den Probanden durch eine Suchoperation "aufzulösen" ist (vgl. Abbildung 12).

Alle der weiter oben angeführten "Regeln (RE)" können in Kombination mit sich selbst oder mit einer anderen RE auf eine der drei oben definierten Arten und Weisen realisiert werden. Dabei wird angenommen, daß die bei IC und IR zusätzlich erforderlichen "Suchoperationen" die Aufgabenschwierigkeit erhöhen.

7.3.6 EINE EXPLORATIVE DATENANALYSE AUF DER GRUNDLAGE DES REVIDIERTEN OPERATIONSINVENTARS

Die Beantwortung der Frage, ob auf der Grundlage hier definierter 'Realisierungsmodi', 'geltender Regeln' und 'Richtungen der

Regelgeltung' HOMOGENE Items konstruierbar sind und ob für alle Operationen interpretierbare Schwierigkeitsschätzungen gewonnen werden können, bleibt einer empirischen Evaluation mit entsprechend konstruierten Items überlassen. Um aber einen ersten Eindruck zu gewinnen, ob und in welcher Weise sich der Wegfall des Faktors 'Materialeigenschaften' und der Richtung der Regelgeltung 'waagerecht und senkrecht' auf die Modellanpassung an einen bereits analysierten Datensatz (vgl. HORNKE und HABON, 1982) auswirkt, wurden die verbliebenen Operationsparameter erneut geschätzt und mit den Schätzungen des revidierten Ansatzes (ALTREGELN) verglichen. Tabelle 2 zeigt, daß die Schätzungen im Faktor 'Richtung der Regelgeltung (RR)' weitgehend konstant bleiben, im Faktor 'geltende Regel (RE)' weist die Operation 'Variieren' nun aber eine betragsmäßig größere Schwierigkeit auf als zuvor.

		ALTREGELN	REVISION
ME	Fo	0.00	--
	M	0.24	--
	R	0.55	--
RE	I	-0.75	-0.75
	V	0.42	0.68
	A	0.00	0.00
RR	w/s	0.20	--
	w	0.00	0.00
	s	1.16	1.12

Tabelle 2:
Vergeich der Operationsparameterschätzungen für das ursprüngliche und das revidierte Modell

Da die Operation 'Variieren (V)' in dem analysierten Aufgabenmaterial häufig mit der nun nicht mehr berücksichtigten Materialeigenschaft 'räumliche Anordnung (R)' gekoppelt war, scheint die Annahme gerechtfertigt, daß 'Variieren' nunmehr auch spezifische schwierigkeitserzeugende Elemente der früheren Operationskombination 'räumliche Anordnung variieren' ausdrückt. Bemerkenswert ist, daß diese Operationskombination im Gegensatz zu anderen Kombinationen mit der Regel 'Variieren' fast ausschließlich mit bestimmten Repräsentanten (Strichen, Pfeilen etc.) realisiert wurden. Das 'Variieren' solcher "offener Gestalten" (Vo) soll im folgenden von dem 'Variieren "geschlossener Gestalten"' (Vierecke, Dreiecke, Kreise etc.) unterschieden werden (Vg).

Differenziert man in der Strukturmatrix nach Vo und Vg und schätzt ihre Schwierigkeitsbeiträge, so wird deutlich, daß Vg etwa die Schwierigkeit von 'Variieren' unter 'ALTREGELN' annimmt, Vo aber einen eigenständigen und beinahe doppelt so großen Schwierigkeitsbeitrag liefert. Abbildung 15 zeigt ein Item des APM, in dem beide Formen des 'Variieren' realisiert sind.

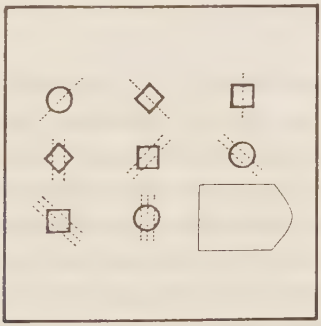


Abbildung 15 :
Beispielaufgabe aus dem APM von RAVEN (1962),
in dem beide Variationsformen realisiert sind.

Die mit der Variation offener Gestalten einhergehenden Raumlageveränderungen der Repräsentanten scheinen diese Operationen gegenüber Vg zu erschweren. Es ist demnach sinnvoll die Konstruktionsregel 'Variieren' folgendermaßen aufzuspalten:

Je Zeile und/oder Spalte des Schemas

--- variiert die Reihenfolge von drei eindimensionalen Repräsentanten ("offene Gestalten").
- VARIIEREN OFFENER GESTALTEN (Vo)

--- variiert die Reihenfolge von drei zweidimensionalen Repräsentanten.
- VARIIEREN GESCHLOSSENER GESTALTEN (Vg)

7.4 EIN RATIONAL ZUR KONSTRUKTION DER WAHLANTWORTEN

Den Antwortalternativen als wichtiger Teilaspekt der Zusammensetzung und Lösung von Matrizenaufgaben wurde bisher bei der Konstruktion vergleichsweise wenig Aufmerksamkeit zuteil. Dies ist insofern verwunderlich, als schon lange bekannt ist, daß Wahlantworten einen moderierenden Einfluß auf das Aufgabenlösungsverhalten ausüben.

So stellte beispielsweise SIEGEL (1963) geschlechts- und altersspezifische, jedoch vom Rohwert unabhängige Bevorzugungen bestimmter Falschantworten bei den Standard Progressive Matrices (SPM) von RAVEN fest. JACOBS und VANDEVENTER (1970) kategorisierten die Falschantworten der SPM als "superior" (teilweise richtig) und "other". Aus der Beobachtung, daß bei 9 von 18 Items die "superior"-Antworten positive Trennschärfen aufweisen, ziehen sie den Schluß, daß Falschbeantwortungen zusätzliche diagnostische Information erbringen.

In der testtheoretischen Literatur finden sich verschiedene Methoden, Informationen aus Falschbeantwortungen von Multiple-Choise Items zu berücksichtigen. Während einige Autoren versuchen mit Hilfe von Expertenratings ALLE Itembeantwortungen gewichtet

in den Testscore einzubeziehen (HAMBLETON, ROBERTS und TRAUB, 1970; PATNAIK und TRAUB, 1973), schlagen andere Autoren eine empirische Gewichtung der Antwortoptionen vor (HENDRICKSON, 1971; REILLY und JACKSON, 1973). Wieder andere fordern die Probanden auf, nicht einfach eine Antwort auszuwählen, sondern die Wahrscheinlichkeit der Antwortalternativen einzuschätzen (HAMBLETON et al., 1970).

Neuere testtheoretische Entwicklungen gestatten es, Informationen aus Falschbeantwortungen direkt in die Schätzung eines oder mehrerer latenter Personenparameter einzubeziehen (vgl. z.B. FISCHER (1974) zu mehrkategorialen Verallgemeinerungen des RM). BOCK (1972) verallgemeinerte das zweiparametrige Modell von BIRNBAUM (1968) zu einem "multiple category model", das die Wahrscheinlichkeit spezifiziert, mit der eine Person der Fähigkeit θ eine der Antwortalternativen k(j) eines Items j mit insgesamt m(j) Kategorien wählt.

$$(43) \qquad p_{jK}(\theta) = \frac{\exp[z_{jK}(\theta)]}{\sum_{h=1}^{m} \exp[z_{jh}(\theta)]}$$

$$\text{mit } z_{jh}(\theta) = c_{jh} + a_{jh}\theta;\ h = 1,2,3 \ldots, K, \ldots m$$

und c(jh), a(jh) als Parameter, die die RELATIVE Schwierigkeit einer Kategorie bzw. die Regression der "Kategorienantwort" auf die latente Fähigkeit (slope-Parameter) charakterisieren (vgl. BOCK, 1972).

In einer Antwortkategorie können entsprechend theoretischer Vorüberlegungen ggfs. mehrere multiple-choice Alternativen zusammengefaßt werden. THISSEN (1976), der Items der CPM und SPM mit dem BIRNBAUM-Modell und dem mehrkategoriellen Modell von BOCK (1972) untersuchte, faßte die Itembeantwortungen in folgende vier Kategorien zusammen: (1) die Richtigantwort, (2) die häufigste Falschantwort, (3) die zweithäufigste Falschantwort und (4) alle

anderen Antworten. CHI**2-Anpassungskontrollen zu beiden Modellen signalisierten signifikante Modellabweichungen für das BIRNBAUM-Modell und befriedigende Konformität der Daten mit dem BOCK-Ansatz. Eine Analyse der Testinformationsfunktion (vgl. SAMEJIMA, 1969) unter beiden Bedingungen, verdeutlicht die Überlegenheit des mehrkategorialen Modells in der unteren Hälfte der Personenparameterverteilung.

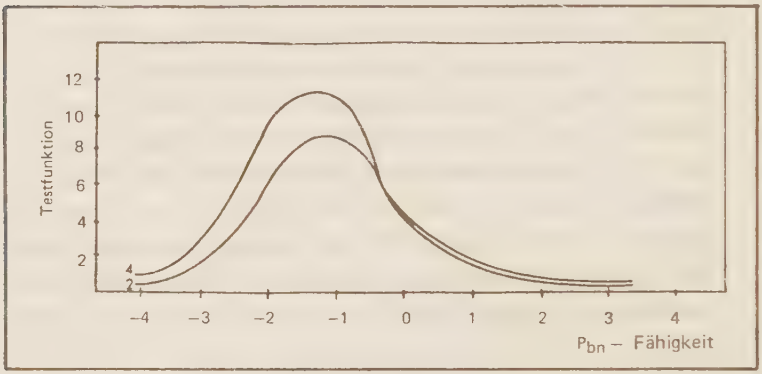


Abbildung 16 :
Testinformationsfunktionen für binäre (2)
und multiple category (4) – Scores,
aus THISSEN (1976)

Gerade, wenn es darum geht, die Personenparameterschätzungen in unteren Skalenbereichen zu präzisieren, ohne den Test zu verlängern, erscheint es gewinnbringend, Zusatzinformationen aus Falschantworten zu berücksichtigen. Vorteile sind besonders dann zu erwarten, wenn insgesamt viele Falschantworten vorliegen, d.h. der Test für die untersuchten Personen eher zu schwer als zu leicht ist.

> "In fact, something may be salvaged from multiple choice tests which were inadvertantly much to difficult for the examinees by a multiple category scoring which would than be based primarily on wrong responses" (THISSEN, 1976, S.213).

Die beobachteten Itemkategoriencharakteristiken des BOCK-Modells für die untersuchten Aufgaben der SPM zeigen eine deutliche Korrespondenz zwischen dem Antwortverhalten der Probanden und den verschiedenen Ähnlichkeitsabstufungen der Distraktoren bezüglich der richtigen Antworten, und zwar etwa in dem Sinne, daß ein Pb mit höherer Fähigkeitsausprägung tendentiell eher einen "superior"-Distraktor wählt.

Hieraus ergeben sich interessante Konsequenzen für eine an Regeln/ bzw. Operationen orientierte Konstruktion von Matrizenaufgaben: sollte es gelingen, Distraktoren zu konstruieren, die der Richtigantwort in dem Sinne ähnlich sind, daß sie bezüglich Art und Zahl erforderlicher Operationen a priori exakt beschreibbare TEILlösungen repräsentieren, so könnten diese Teillösungen zum inhaltlich bestimmten Einteilungskriterium für die Kategorienbildung im "BOCK-Modell" gemacht werden. Eine inhaltlich begründete Kategorienbildung hat gegenüber dem Vorgehen von THISSEN - der die Distraktoren ja bezüglich ihrer Auftretenshäufigkeit gruppierte (s.o.)- den Vorteil, daß die aus relativen "Falsch"-antworten gewonnene zusätzliche diagnostische Information in die Interpretation des durch die Regeln/ Operationen definierten Latent Trait eingeordnet werden kann, d.h. sie ist "konstruktvalide" im o.a. Sinn.

Eine rationale Distraktorenkonstruktion muß so angelegt sein, daß die Konstruktionsrationale nicht von den Probanden erkannt und zur Lösungsfindung genützt werden können. So zeigen WHITE und ZAMARELLI (1981) u.a. am Beispiel von 12 Items der APM, daß durch Beachtung einfacher "convergence principles" die "Antwortsets", bestehend aus Richtigantworten und Distraktoren, auf Untermengen reduziert werden können, die im Extremfall nur ein Element, nämlich die richtige Lösung enthalten. Eine Untersuchung, in der lediglich die Wahlantworten des CFT (CATTELL und CATTELL, 1973) vorgelegt wurden, zeigte deutlich, das Probanden allein aufgrund der in den Antworten vorfindbaren Konvergenzprinzipien, die

richtige Lösung ÜBERZUFÄLLIG häufig identifizieren.

Die einzige sichtbare Möglichkeit solche Schwierigkeiten zu umgehen ist, nach Ansicht der Autoren (S.27),

> "to draw up sets of distractors for items in such a way that the answer sets contain no information as to which of their elements are the correct answers."

GUTTMAN und SCHLESINGER (1967) weisen auf den Einfluß hin, den Distraktoren auf die Aufgabenschwierigkeit haben:

> "Distractors are an integral part of the stimulus situation and determine the degree of difficulty of an item in conjunction with the question part. The question part of an item will not correspond to any definite level of difficulty when construction of distractors is haphazard" (S. 570).

Die Autoren schlagen daher vor, Distraktoren rational durch "facet designs" zu konstruieren (vgl. auch ROID und HALADYNA, 1982; HORNKE und HABON, 1983). Bezogen auf Matrizenitems bedeutet dies eine systematische Variation struktureller und phänomenologischer Merkmale des ItemSCHEMAS in den Distraktoren. Eine Untersuchung mit entsprechend konstruierten Aufgaben bestätigte die Hypothese, daß die Attraktivität eines Distraktors in direkter Beziehung zur Ähnlichkeit mit der Richtigantwort steht, wobei GUTTMAN und SCHLESINGER (1967) "Ähnlichkeit" A PRIORI durch die Anzahl der "facets" definierten, die der Distraktor mit der richtigen Lösung gemeinsamt hat. Ersetzt man hier den Begriff der "facets" durch den Regel-/ Operationsbegriff, so ist damit die bereits oben geforderte konstruktvalide Distraktorenkonstruktion angesprochen.

Welche Probleme sich dabei im vorliegenden Fall stellen, soll anhand eines Beispiels demonstriert werden.

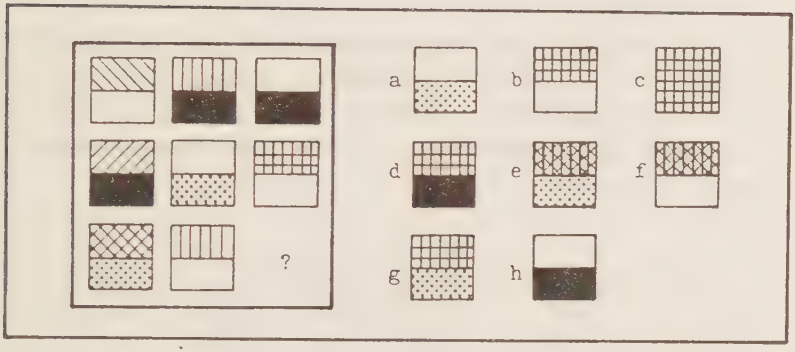


Abbildung 17 :
Beispielaufgabe

Diese Aufgabe setzt sich aus zwei Komponenten zusammen, die in jeder Figur durch einen waagrechten Strich getrennt werden. Angenommen, ein Pb versucht die Aufgabe zu lösen, indem er zunächst nach einer Regel für die Komponenten in der oberen Hälfte der Figuren sucht. Durch 'Addition' des Musters in den Spalten des Schemas stellt er fest, das der richtige Lösungsvorschlag in der oberen Hälfte eine Gitterstruktur aufweist. Ein Blick zu den Antwortalternativen zeigt ihm, daß gleich 4 Wahlantworten (b,c,d und g) entsprechend aufgebaut sind und wiederum 3 davon in der unteren Hälfte alle denkbaren Lösungen für die zweite Komponente enthalten. Der Pb kann die Aufgabe also nur lösen, indem er auch für diese Komponente im Figurenschema nach einer Aufbauregel sucht. Das Beispiel verdeutlicht, daß durch eine TEILlösung der Aufgabe die Zahl potentieller "Richtigantworten" eingeschränkt wird. Das Ausmaß der Einschränkung hängt davon ab, wieviele richtige Teillösungen sich für beide Komponenten unter den Antwortvorschlägen befinden. Während die Beispielaufgabe den "idealen" Fall minimaler Einschränkung darstellt, führt im ungünstigsten Fall -bei gleichartigen Lösungsverhalten- bereits eine Teillösung zur richtigen Aufgabenbeantwortung (s. WHITE und

ZAMARELLY, 1981). Da viele Pbn bei der Aufgabenlösung nicht nur das Figurenschema, sondern auch die Wahlantworten analysieren (vgl. dazu PUTZ-OSTERLOH, 1981, S.145ff.), sollte für eine geeignete Verteilung "richtiger" und "falscher" Komponenten in den Antwortvorschlägen gesorgt werden. Zur Objektivierung der Distraktorenkonstruktion wird daher folgender Verteilungsschlüssel (DS) vorgeschlagen.

	r		f - gl. Repr.-Paare		f - einz. Repr.	
DS-Nr	1.K.	2.K.	1.K.	2.K.	1.K.	2.K.
1	2	3	2	2	1	0
2	2	2	2	2	1	1
3	1	2	2	2	2	1

Tabelle 3:
Schlüssel zur Konstruktion der Wahlantworten

Richtiglösungen und Teillösungen wurden in der Beispielaufgabe entsprechend DS-Nr. 1 verteilt. Die Zuweisung der richtigen Gesamtlösung auf die Position (g) erfolgte nach einem Zufallsprinzip. Für die ERSTE Komponente (1.K.-UNTERE Hälfte der Figuren) wurden zwei richtige (r) Teillösungen zufällig auf die Positionen (e,a) verteilt, zwei Paare jeweils gleicher falscher (f), aber "möglicher", d.h. im SCHEMA vorkommender Repräsentanten auf die Positionen (d,h und b,f), sowie ein einzelner falscher und "unmöglicher" Repräsentant auf die Position c. Für die zweite Komponente (2.K.-OBERE Hälfte der Figuren) wurde analog verfahren.

Eine derartige regel-/ operationsorientierte Zusammenstellung der Wahlantworten soll eine gezielte Teillösungs-/ Fehleranalyse möglich machen. Abgesehen von der bei diesem Aufgabenformat wohl nicht auszuschließenden Möglichkeit, daß Pbn durch Raten zur

richtigen oder zumindest im o.a. Sinne "teilweise richtigen" Lösung kommen, können bei rationaler Distraktorenkonstruktion fehlerhafte Strategien identifiziert und in Operationsbegriffen benannt werden. Hierzu ein Beispiel: Entfällt die Antwort eines Pb bei obiger Beispielaufgabe auf die Alternative 'e', kann angenommen werden, daß er die spaltenweise (senkrechte (s)) Variation des Musters (Vg) in der unteren Hälfte der Figuren erkannt und durchgeführt hat. Die zusätzliche senkrechte 'Addition (A)' der Linien in den oberen Hälften hätte ihn zur richtigen Lösung 'g' geführt. Distraktor 'e' legt hingegen die Hypothese nahe, daß er anstelle der 'richtigen' Operationskombination ('A' und 's') die Folge ('A' und 'w') auf die dritte Zeile des Schemas angewendet hat. Es wäre zu untersuchen, inwieweit sich Pbn hinsichtlich solcher Lösungs-/ Fehlerstrategien systematisch unterscheiden und wie dies ggfs. diagnostisch nutzbar gemacht werden kann. Möglichkeiten hierzu scheinen mehrkategorielle logistische Testmodelle zu bieten (s.o.: THISSEN, 1976).

8. DIE KONSTRUKTION DES ITEMPOOLS

Bei Beschränkung auf Items mit je ZWEI separiert oder integriert dargebotenen, zeilen- und/oder spaltenbezogen geltenden Regeln sowie VOLLSTÄNDIGER Kombination aller Operationen läßt das Konstruktionsrational für figurale Denkaufgaben prinzipiell die Konstruktion von 648(!) sturkturVERSCHIEDENEN (=nicht-parallelen) Aufgaben zu.

Die Strukturmatrix der Aufgabengrundmenge stellt den Ausgangspunkt für die Konstruktion dar. Sie gibt Aufschluß über die Aufbauregeln und deren Kombination. Da in Anlehnung an vorgängige Itemkonstruktionen nur EINE Regel-/ Operationswiederholung pro Aufgabe zugelassen werden soll, ist diese Matrix ausschließlich aus den Operationshäufigkeiten f(gj)=(0,1,2) aufgebaut. Die Matrix spezifiziert dann 144 Items, in denen jeweils eine geltende Regel zweimal anzuwenden ist und 504 Aufgaben, in denen zwei verschiedene Regeln in Kombination auftreten.

Um die Konstruktion weiter zu objektivieren, wurde die von HABON (1981) zusammengestellte Liste figürlicher Repräsentanten erweitert, die Einzelfiguren indiziert und die Konstrukteure auf diesen Symbolvorrat bei der Aufgabenerstellung verpflichtet. Die "Repräsentantenliste" (Abb.18, s. nächste Seite) soll möglichen Iteminhomogenitäten vorbeugen, die durch allzu "phantasievolle" graphische Repräsentationen entstehen könnten. Eine derartige Beschränkung scheint darüber hinaus auch deshalb sinnvoll, da beispielsweise bei einer denkbaren automatisierten Itemkonstruktion, wohl zwangsläufig auf eine geringe Zahl zuvor festgelegter Figurenelemente zurückgegriffen werden müßte. Darüber hinaus bietet die zusätzliche Beschreibung von Items durch die Indizes der verwendeten Repräsentanten Möglichkeiten Inhomogenitäten oder Itemschwierigkeiten, die nicht allein durch die angenommenen Operationen erklärbar sind, systematisch auf das phänomenologische Erscheinungsbild der Aufgaben zurückzuführen.

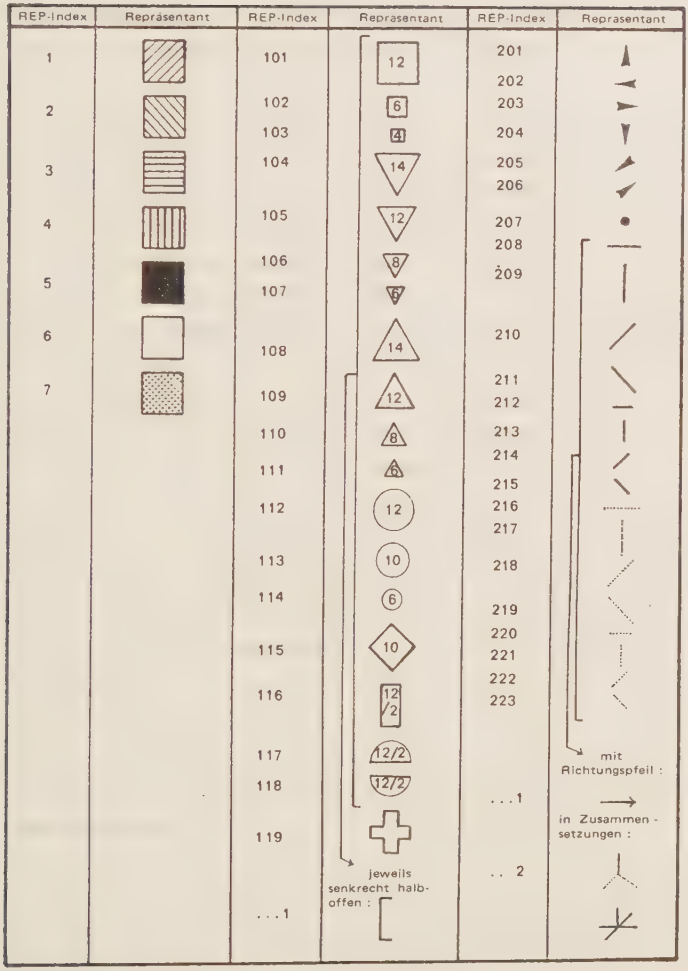


Abbildung 18 :
Liste der Repräsentanten (verkleinert)

Die Konstruktionsarbeiten führten mehrere studentische Mitarbeiter nach entsprechender Einweisung unter Supervision durch. Die Aufgaben wurden zunächst auf Vordrucken entworfen, auf denen auch die zufällig zugewiesene Schlüsselnummer des Distraktorenrationals und die Indizes der verwendeten Repräsentanten eingetragen wurden. Im Anschluß an diese Vorarbeiten wurden die Vorlagen ausgetauscht und überprüft ob sie der Konstruktionsvorschrift entsprechen. Es zeigte sich, daß der Modus "integrierte Komponenten (IC)" für Kombinationen der Operation "Identität (I)" mit anderen Operationen oder mit sich selbst nicht realsiert werden kann, wenn die "Richtungen der Regelgeltung" identisch sind. "IC" sollen ja nur dann definiert sein, wenn relevante Merkmale eines Figurenelements NICHT-konstante, irrelevante Merkmale des anderen Figurenelements darstellen. Ebenso wurde auf die Konstruktion von Items verzichtet, in denen gleichgerichtete "Identitäten" mit integrierten Repräsentanten (IR) realisiert werden sollten. Eine solche Darstellung ist ebenfalls definitorisch ausgeschlossen, da entsprechend konstruierte und aus zwei Komponenten zusammengesetzt gedachte Figuren als einzige Gesamtfigur aufgefaßt werden können. Im Gegensatz zu anderen Operationskombinationen wäre hier die intendierte Suchoperation zur Auflösung der "geschlossenen Gestalt" nicht erforderlich, um die Aufgaben erfolgreich zu bearbeiten.

Die Itemgrundmenge reduzierte sich somit um insgesamt 32 auf 616 verschiedene Aufgaben, die mit standardisierten Zeichenmaterialien auf DIN A 4-Vordrucke übertragen wurden. Diese Vorlagen wurden anschließend fotografisch reproduziert, auf Karteikartenformat vergrößert und schließlich zu Druckvorlagen zusammengestellt. Dieses Verfahren sicherte eine hohe Druckqualität, die nach bisherigen Erfahrungen gerade bei figuralem Aufgabenmaterial bedeutsam ist.

9. DIE VERNETZUNG DER TESTVERSIONEN

In der traditionellen Testkonstruktion tritt die Frage, ob eine größere Aufgabenmenge ein und dasselbe Merkmal erfaßt, hauptsächlich im Zusammenhang mit Paralleltestkonstruktionen auf. Dabei gelten in der klassischen Testtheorie Tests als äquivalent, wenn sich die Verteilungen der Rohwerte beider Testformen entsprechen. Ist dies nicht der Fall, können beispielsweise zur Angleichung der Verteilungen die Itemschwierigkeitsrangreihen der Tests verändert oder bei Mittelwertsunterschieden leichte oder schwere Items der einen oder anderen Testform hinzugefügt werden. Sind auch die Rohwertvarianzen signifikant verschieden, und verzichtet man aus ökonomischen Gründen auf Testrevisionen oder unabhängige Testnormierungen, so wird in der Regel versucht, die Tests mittels der Äquiperzentil-Methode einander anzugleichen (vgl. LIENERT, 1969(3), S.358ff.). Dieses Verfahren erfordert aber umfangreiche Personenstichproben, in denen das Merkmal so verteilt ist, daß alle belangvollen Score- Perzentil- Verknüpfungen adäquat vertreten sind. Um die Zahl der von einem einzelnen Probanden zu bearbeitenden Items dennoch in tolerablen Grenzen zu halten, schlägt LIENERT (1969(3)) konsequenterweise vor, die Testversionen unabhängigen Pbn-Zufallsstichproben vorzulegen. Im Zusammenhang mit der Evaluation großer Itemsammlungen in den USA wurde dieser Grundgedanke in den 70er Jahren auch auf die Zufallsauswahl von Items, genauer, von Item-Pbn-Stichproben ausgeweitet (vgl. SHOEMAKER, 1973 und HORNKE, 1976b zum "multiple-matrix-sampling").

Mit der durch das RM gegebenen Möglichkeit, spezifisch objektive -von der jeweiligen Zusammensetzung der Personenstichprobe unabhängige- Aufgabenschwierigkeitsschätzungen zu gewinnen, erübrigen sich solch komplexe und schwer kontrollierbaren Stichprobenpläne (vgl. CHOPPIN, 1968, 1976; WILMOTT und FOWLES, 1974, S.46-51). Die aus Antwortdaten beliebiger unabhängiger Pbn-Gruppen geschätzten Parameter zweier gleicher Items sind bei hinreichender

Modellkonformität statistisch äquivalent bis auf eine Translationskonstante. Unterstellt man einmal Homogenität und meßfehlerfreie Schwierigkeitsschätzungen für ALLE Items der beiden fraglichen Tests, so überführt diese Konstante alle Itemparameter auf eine gemeinsame Skala. Umfangreiche Itempools sind nun auch kalibrierbar, wenn die Aufgaben auf verschiedene Tests verteilt und diese von unabhängigen Pbn-Gruppen bearbeitet werden. Voraussetzung ist lediglich, daß die Tests nicht ausschließlich verschiedene ("tag"-)Items enthalten, sondern das mindestens jeweils zwei Tests durch gemeinsame Aufgaben in einer Linkstruktur miteinander verbunden sind, sodaß Translationskonstanten bestimmbar sind. In der Praxis kann die Homogenität der Items und ihre meßfehlerfreie Schätzung nicht vorausgesetzt werden. Ob alle auf mehrere Tests verteilten Aufgaben homogen im Sinne des RM sind, ist kontrollierbar, indem man die Schätzwerte gleicher "link"-Items mittels graphischer und/oder statistischer Modellkontrollen hinsichtlich ihrer statistischen Äquivalenz überprüft (vgl. dazu ANDERSON, 1973; WRIGHT und STONE, 1979).

Da Itemparameterschätzung fehlerbehaftet sind, werden die Translationskonstanten geschätzt. Werden zwei durch K Linkitems verbundene Tests A und B verschiedenen "Stichproben" von N Personen vorgegeben und bezeichnen b(A) und b(B) die Schwierigkeitsschätzungen der Aufgaben in jedem Test mit einem ungefähren Standardfehler von 2.5/N**1/2, so beträgt die Konstante G zur Überführung der Schätzwerte ("horizontal equating") in Test B auf die Skala des Tets A:

$$(44) \qquad G_{AB} = \sum_{g}^{K} (b_{gA} - b_{gB})/K$$

wobei der Standardfehler der Differenz-Statistik etwa 3.5/(NK)**1/2 Logits beträgt (vgl. WRIGHT und STONE, 1979, S.96ff.). WRIGHT und STONE (1979) schlagen verschiedene Möglichkeiten vor, die Items eines umfangreichen Pools in diesem Sinne

zu verketten bzw. zu vernetzen (vgl. S.98ff.).

Zur Kalibrierung des Matrizenitempools soll eine relativ leicht handhabbare und überschaubare Verkettungsstruktur konstruiert werden, die erstmals auch Aufbaumerkmale der Items explizit berücksichtigt und folgende ALLGEMEINE Struktur aufweist:

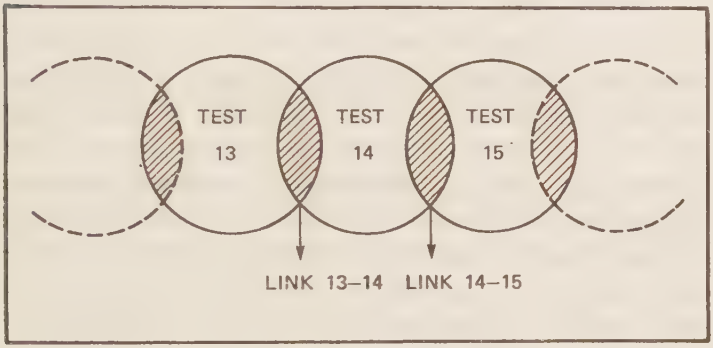


Abbildung 19:
Ausschnitt der verwendeten Linkstruktur

Die Tests sollen insgesamt 24 Aufgaben, jeder Link, der jeweils zwei Tests verbindet, je 6 Aufgaben enthalten. Um auf diese Weise 35(!) Tests zu verketten sind (34x6=) 204 Linkitems erforderlich. Die Tests '1' und '35' sind mit je einem weiteren Test ('2' bzw. '34') verkettet und enthalten deshalb (24-6=) 18 Tagitems. Die übrigen Tests (2-34) sind über 12 Linkaufgaben mit jeweils zwei Tests verbunden und enthalten zusätzlich 12 weitere Items. Da der Bedarf an (33x12)+(2x18)= 432 Tagitems aus den verbliebenen 412 verschiedenen Aufgaben (-Strukturen) des Pools nicht gedeckt werden konnte, wurden zu zwei Linkitem- und 16 Tagitemstrukturen je zwei (phänomenologisch) verschiedene Aufgaben konstruiert. Eine Tagitemstruktur wurde dreifach verwendet.

Liegen keinerlei Hypothesen über den strukturellen Aufbau oder die Schwierigkeit der Aufgaben eines Pools vor, wird man ihnen

u.U. durch ein Zufallsverfahren einen bestimmten Platz in der 'Kette' zuweisen. Der Gefahr, Tests mit wenig differenzierten, einseitigen Operationsstrukturen oder generell zu "leichten" oder zu "schweren" Aufgaben zusammenzustellen, soll hier begegenet werden, indem alle Aufgaben Strukturmerkmale gleichmäßig und systematisch auf die Test verteilt werden.

Da die Dimensionalität des Pools erst über die aus jeweils 6 Items bestehenden Links prüfbar wird, ist ihrer Zusammensetzung besondere Aufmerksamkeit zu widmen. Die unter den Realisierungsmodi 'IC' und 'IR' zusätzlich erforderlichen Suchoperationen stellen vermutlich noch am ehesten Quellen möglicher Iteminhomogenitäten dar (vgl. PUTZ-OSTERLOH, 1981, S.170). Um einem evtl. völligen "Bruch" der Verkettungsstruktur vorzubeugen, scheint es daher angebracht, die Realisierungsmodi 'SC', 'IC' und 'IR' gleichhäufig, also immer zweifach, in den Link einzubringen. Mit den "Richtungen der Regelgeltung" (w, s, w/s) wird analog verfahren (vgl. den unten dargestellten Ausschnitt der Linkstruktur, Abbildung 20). Der Link befindet sich dabei nicht geblockt an irgendeiner (möglicherweise störanfälligen) Position des Tests, sondern verteilt sich über zuvor bestimmte Positionen (s.u.).

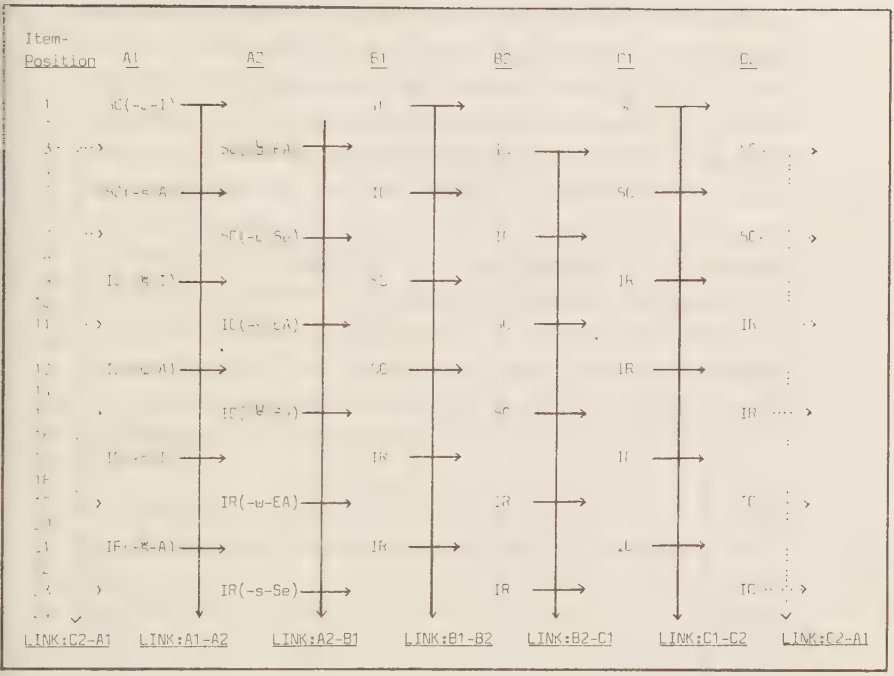


Abbildung 20:

Verteilung der Strukturmerkmale über die Items

Die beiden in einem Test befindlichen Link-Blocks greifen ineinander und bilden ein Gerüst, in das die Tagitems nach weiter unten spezifizierter Vorschrift eingefügt werden. Innerhalb der Link-Blocks sind die Realisierungsmodi auf drei verschiedene Arten A, B und C angeordnet (vgl. Abb. 20). Ausgehend von der Hypothese, daß Items mit 'SC' tendentiell leichter sind als solche mit 'IC' und diese wieder leichter als Aufgaben, in denen "geschlossene Gestalten" (IR) aufzulösen sind, sind im Blocktyp A 'SC'- vor 'IC'- vor 'IR'- Items angeordnet. Um evtl. Strategieinduktionen, die durch massiertes Auftreten gleicher Realisierungsmodi in aufeinanderfolgenden Items hervorgerufen werden,

kontrollieren zu können, vertauschen in den Blocktypen B und C die Modi 'SC' und 'IC' bzw. 'IC' und 'IR' ihre Positionen. Die restlichen (eingeklammerten) Itemstrukturen bleiben jedoch unverändert. Da sich ein Blocktyp immer über zwei Tests erstreckt (A1/A2...), haben die Linkitems jedes siebten Tests bezüglich mindestens einer der beiden Itemkomponenten einen identischen strukturellen Aufbau.

Die Verteilung der "tags" auf die verbleibenden 12 bzw. 18 Positionen ist nun so organisiert, daß 'SC', 'IC' und 'IR' gleichhäufig und über die Tag-Positionen abwechselnd verteilt sind.

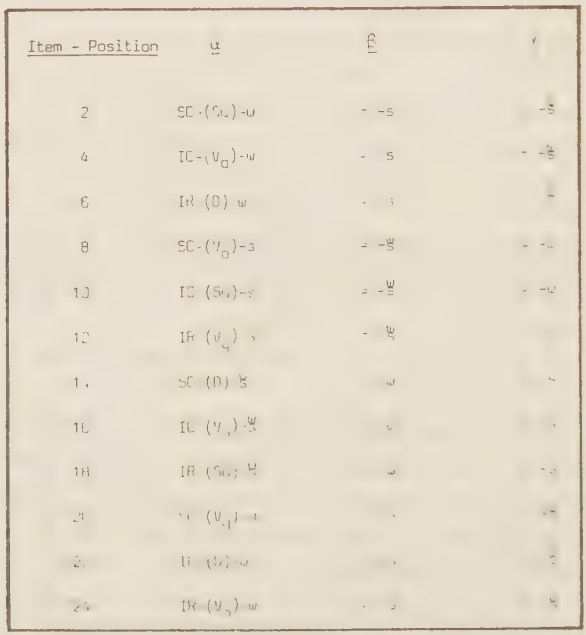


Abbildung 21 :
Verteilung der Strukturmerkmale über die Tagitems.

Anordnung und Häufigkeit der 'Richtungen der Regelgeltung' (w, s, w/s) werden dabei jeweils über drei Tests (α β γ) systematisch variiert, so daß die Tagitems jedes vierten Tests partiell (d.h. für je eine Itemkomponente) identische Strukturen aufweisen. Jeder siebte Tests ist daher bezüglich mindestens einer Komponente von Links und Tags strukturell parallel. Die (eingeklammerten) "geltenden Regeln" (s. Abb. 21) der Tagitems sollen die der Linkitems (vgl. Abb. 20, dort ebenfalls eingeklammert) wann immer das möglich ist, so ergänzen, daß jede Operation in jedem Test insgesamt mindestens dreimal vorkommt, wobei in aufeinanderfolgenden Aufgaben nicht dieselben "Regeln" gelten sollen. Dieses Verteilungsprinzip konnte allerdings nicht vollständig durchgehalten werden, hauptsächlich, weil einige Struktur- bzw. Operationskombinationen auszuschließen waren (s. Kp. 8).

Die systematische Verteilung von Strukturen über die Testkette bezieht sich immer nur auf eine der beiden Komponenten (bestehend aus je einer Regel und deren Richtung), aus denen sich eine Aufgabe zusammensetzt. Bei der Zusammenstellung der Testhefte (vgl. Kp.10) wurden dem o.a. Verteilungsschlüssel entsprechend Aufgaben aus dem Pool "gezogen", wobei sich die zweite, hinzukommende Itemkomponente jeweils zufällig ergab. Da sich keine (!) vollständig strukturparallelen Aufgaben in der Itemgrundmenge befinden, existieren gemäß dem Verteilungsschlüssel der Verkettungsstruktur auch keine vollständig strukturparallelen Test.

Anhang 1 ist zu entnehmen, wie sich die drei Realisierungsmodi, 8 Regeln und drei Richtungen der Regelgeltung, sowie die Distraktorenschlüssel und richtig Antwortpositionen über die 35 Tests verteilen.

10. TESTMATERIAL UND DATENERHEBUNG

In jeder der 35 Versionen wurden 24 Matrizenaufgaben auf 6 Seiten verteilt und zusammen mit je 15 ebenfalls regelgeleitet konstruierten Wortanalogieaufgaben zu Testheften zusammengestellt.

Nach einem Deckblatt, der Instruktion für die Wortanalogieitems und den 15 Analogien folgt ein einfaches, kommentiertes Einführungsbeispiel, das mit dem allgemeinen Lösungsprinzip von Matrizenaufgaben vertraut macht. Außerdem werden einige allgemeine Hinweise zu Bearbeitungsstrategien und dem gewünschten Korrekturmodus gegeben. Auf der nächsten Seite folgen drei Übungsbeispiele mit Angabe der richtigen Lösungen. Zwei Übungsaufgaben, wie auch das Einführungsbeispiel auf der Vorseite, wurden dem WIENER-MATRIZENTEST (FORMANN und PISWANGER, 1979) entnommen. Bei selbständiger Bearbeitung der drei Aufgaben sind - Gültigkeit der Hypothesen zur Aufgabenstruktur vorausgesetzt- die Operationen Vg, SE, A und 'senkrechte Regelgeltung erkennen (s)' jeweils einmal auszuführen. Da hier lediglich das durch die allgemeine Aufgabenform bedingte Antwortverhalten, nicht jedoch die Lösungsoperationen geübt werden sollte, handelt es sich bei den drei Übungsaufgaben um vereinfachte Einkomponentenitems. Um sie zu lösen, sind die vermutlich schwierigkeitserhöhenden Suchoperationen IC und IR, die mehrere Komponenten in einem Item wahrnehmungsmäßig differenzieren, nicht erforderlich.

Nachdem die Übungsaufgaben bearbeitet sind, beginnt der Proband selbständig die nachfolgenden 24 Items zu lösen.

Die gesamte Datenerhebung wurde auf Anordnung des Bundesministeriums der Verteidigung vom Bundeswehrverwaltungsamt organisiert und bundesweit in 33 Kreiswehrersatzämtern (KWEA) und deren Wehrpsychologischen Diensten im Frühjahr 1983 durchgeführt.

Da die vorliegende Itemmenge in erster Linie in der routinemäßigen Eingangsdiagnostik der Bundeswehr eingesetzt werden soll, war bezüglich der Stichprobenzusammensetzung eine breite regionale Streuung und vor allem eine Schulbildungsgleichverteilung in allen Teilstichproben, d.h. für jede der 35 Versionen anzustreben. Letzteres ist schon deshalb von Bedeutung, um einerseits über ausreichend viele Datenpunkte zur Schätzung der Parameter leichter und schwerer Aufgaben zu verfügen und andererseits "Schulbildungsabhängigkeit" als mögliche Quelle der Modellungültigkeit kontrollieren und ausschließen zu können. Letztlich sollte die Datenerhebung so geplant und durchgeführt werden, daß von N=200 Pbn ungefähr n(1)=100 Hauptschulabsolventen mit und ohne Abschluß und n(2)=100 Probanden mit Realschulabschluß eine Testversion vorgelegt wird.

Diese Forderung konnte in den meisten Fällen weitgehend erfüllt werden. In der Gesamtstichprobe ist das Verhältnis von "Abiturienten/Realschülern" zu "Hauptschülern" in etwa ausgeglichen (44% (3.259) "Hauptschüler", 56% (4.153) "Abiturienten" und "Realschüler"). Von den insgesamt 7.416 getesteten Pbn sind 470 (6.3%) "Hauptschüler ohne Abschluß", 2.789 (37.6%) "Hauptschüler mit Abschluß", 1.753 (23.6%) "Realschüler" und 2.400 (32.2%) "Abiturienten". Die Stichprobengrößen liegen für die einzelnen Testversionen zwischen n=178 und n=239, im Mittel bei n=212. Da ausschließlich Wehrpflichtige zum Zeitpunkt der routinemäßigen Eignungs- und Verwendungsprüfung (EVP) der Bundeswehr an der Testerprobung teilnahmen, ist die Stichprobe bezüglich der Geschlechts- und Altersvariable homogen. Wehrpflichtige, die aus medizinischen Gründen wehrdienstuntauglich sind, sowie anerkannte Wehrdienstverweigerer nehmen an der EVP nicht teil. Bis auf die genannten Einschränkungen ist die Stichprobe nicht vorselegiert und kann daher als weitgehend repräsentativ für die wehrpflichtige männliche Bevölkerung der Bundesrepublik Deutschland in dieser Jahrgangskohorte betrachtet werden.

Bemerkenswert scheinen einige Angaben der Kreiswehrersatzämter über die je nach KWEA stark differierenden Bearbeitungszeiten für die Matrizentests. Hier könnte gewisser Zweifel bestehen, ob überall die notwendigen Powerbedingungen hergestellt wurden. Auf Einzelheiten wird im Zusammenhang mit der Ergebnisdarstellung näher einzugehen sein. Bedeutsam ist auch der beinahe durchgängige Hinweis auf Abbrecher, die offensichtlich zumindest bezüglich des Matrizentests deutlich überfordert waren. Nach Angaben der Testleiter handelte es sich hierbei hauptsächlich um Pbn mit und ohne Hauptschulabschluß. Da Rateverhalten unter solchen Bedingungen wahrscheinlich ist, wird dieser Aspekt bei Fragen der Modellanpassung besonders zu berücksichtigen sein.

10. ERGEBNISSE - DARSTELLUNG UND DISKUSSION

10.1 GÜLTIGKEITSKONTROLLEN ZUM RM

10.1.1 DIE ZUSAMMENSTELLUNG "HOMOGENER" PBN-STICHPROBEN

Für die Probanden-Fitanalysen wurden aus jedem der 35 Datensätze approximativ Item- und Personenparameterschätzungen nach "PROX" (vgl. Kp.5.1) durchgeführt. Wie die Rohwertverteilungen (vgl. Anhang 2) zeigen, scheint die Annahme annähernd normalverteilter Personenparameter für alle Datensätze haltbar. Weder bei den Personen- noch bei den Itemrandsummen der Tests konnten schiefe Verteilungen beobachtet werden. Deutliche Floor- oder Ceiling-Effekte, die PROX-Schätzungen problematisch werden lassen, liegen nicht vor. Die methodischen Überlegungen zur Kalibrierung eines homogenen Itempools (Kp.5.3) legen die Selektion von Probanden nahe, wenn ihre beobachteten Aufgabenbeantwortungen Zweifel an deren Interpretierbarkeit als Manifestationen einer eindimensionalen latenten Fähigkeit begründen. Erwartungswidrige, bei Gültigkeit des Modells unwahrscheinliche Reaktionen, sprechen zwar nicht grundsätzlich gegen die Meßbarkeit bestimmter Pbn mit diesem Itemmaterial, da das Modell solche Reaktionen durch die Annahme einer logistischen Wahrscheinlichkeitsfunktion bei gegebenen Parametern sogar "vorsieht". Da es hier jedoch zumindest nicht vorrangig um die Messung der latenten Personenfähigkeit, sondern um die vorgelagerte Skalenkonstruktion geht, scheint es angebracht, die Itemparameter aus im Sinne des RM plausiblen Beantwortungen zu schätzen.

Die einseitig zu interpretierenden, geschätzten Personenmisfitstatistiken t(v) sind nur für wenige Tests mit geringer Varianz systematisch um einen zufriedenstellend niedrigen Median bzw. Mittelwert von etwa $-.2<z(v)<.2$ verteilt (vgl. Abbildung 22a).

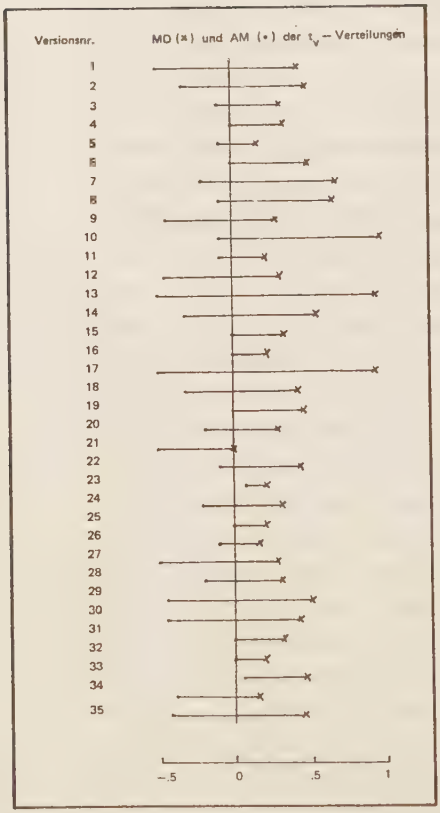
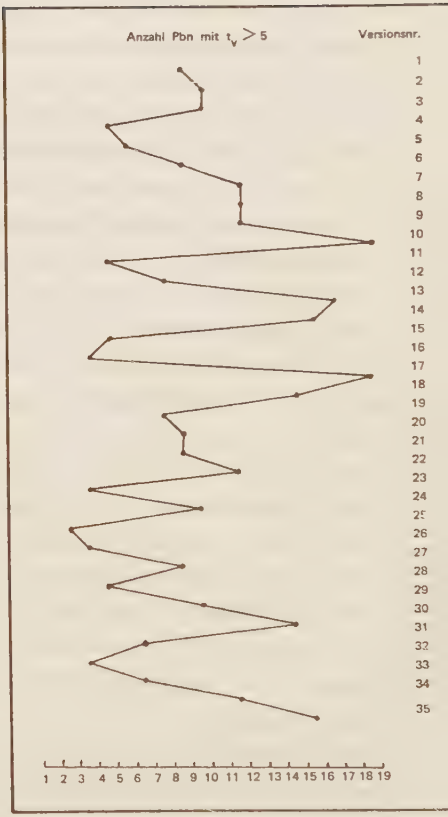


Abbildung 22a:
Lagestatistiken der t_v – Verteilungen



Abbildungen 22b:
Anzahl der Probanden mit $t_v > 5$ in den Tests 1 – 35

In der Mehrzahl der Tests wiesen einige wenige Probanden einen z.T. erheblichen Misfit auf. Dies drückt sich in großen Differenzen zwischen den auch hier niedrigen Medianwerten und den arithmetischen Mittelwerten der entsprechenden t(v)- Verteilungen aus. Besonders deutlich wird dies bei den Tests 10, 13 und 17. Wie der Abbildung 22b zu entnehmen ist, kann das RM bei diesen Tests das Antwortverhalten von mehr als 8% der Pbn nur unzureichend beschreiben (z(v)>5). In 10 Tests fallen hingegen weniger als 5 Pbn (2.5%) durch inkonsistentes Antwortverhalten auf. Die Ursachen dieses stark unterschiedlichen "Stichprobenfits" sind möglicherweise in unterschiedlichen Stichprobenzusammensetzungen, wahrscheinlicher aber in differierenden Instruktions- und vor allem Durchführungsbedingungen zu suchen. Auf diesen Punkt wird im Zusammenhang mit der Itemanalyse noch gesondert einzugehen sein.

Für die später folgenden CML-Schätzungen werden aus den Datensätzen 1-35 insgesamt 302 Antwortvektoren selegiert, die zu einem z(v)>5 führten. Bei Überschreitung dieses kritischen Wertes kann, nach WRIGHT und STONE (1979, S.168ff.), mit einiger Berechtigung davon ausgegangen werden, daß eine Parametrisierung des Pbn-Verhaltens nicht sinnvoll ist. Hingegen können die Statistiken bereits dann einen Wert im schon kritischen Bereich 3<t(v)<5 annehmen, wenn durch Itemmisfit bedingte einzelne hohe Residuen in die Summenstatistik eingehen. Sinnvolle Probandenselektion setzt in einem solchen Fall ein gemischtes Selektionskriterium voraus, das den Pb als Quelle des Misfit hinreichend zuverlässig zu identifizieren vermag. Im vorliegenden Fall soll daher eine Selektionsentscheidung getroffen werden, wenn bei einem Pb mit t(v)>3 mindestens drei "auffällige" Beantwortungen (z(vg)>10) auf Items beobachtet werden, die ihrerseits keinen bedeutsamen Misfit erkennen lassen (t(g)<5). Dieses Mischkriterium trifft auf insgesamt 117 Pbn zu, von denen allerdings etwa die Hälfte (59 Pbn) einen t(v)>5 realisieren. Es sind somit zusätzlich 58 Pbn zu selegieren, die sich, wie in der zusammenfassenden Abbildung 23

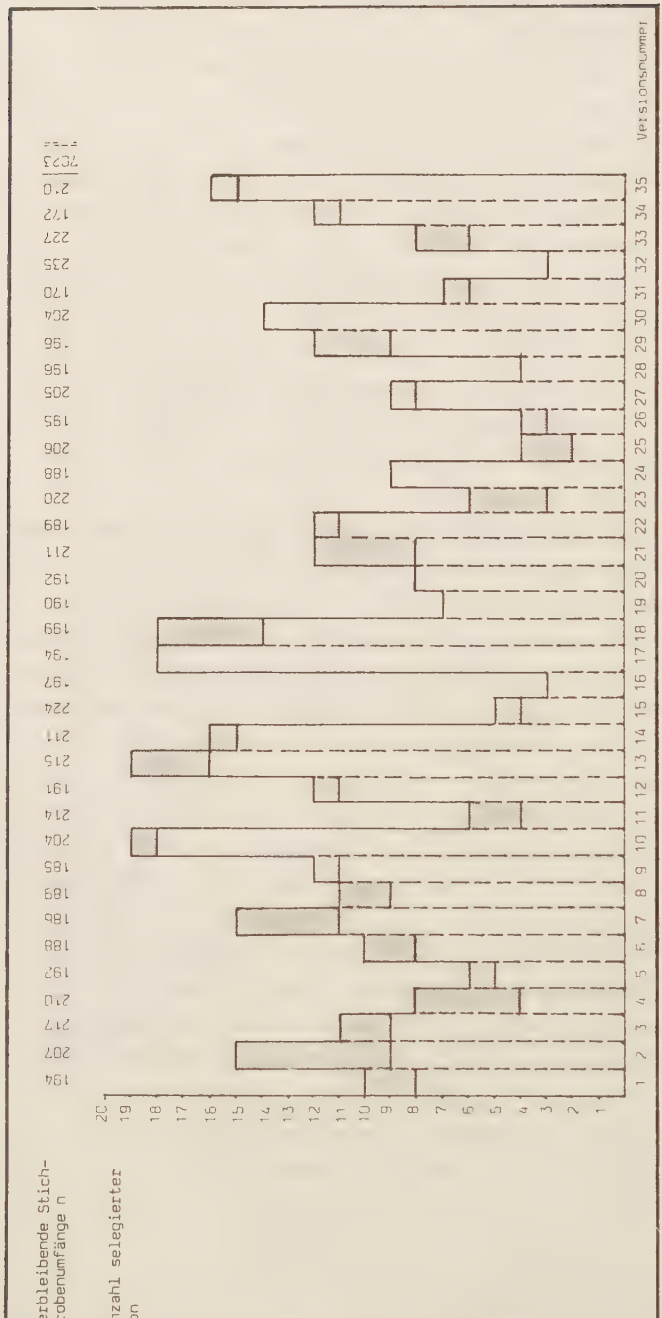


Abbildung 23 :
Anzahl der in den Tests 1 – 35 wegen mangelnden Fits
selegierten Probanden

(durch Schwärzung) dargestellt, auf die 35 Tests verteilen. Die nach letztendlicher Selektion von min. 3 bis max. 19 (im Durchschnitt etwa 10=5%) der Probandenvektoren pro Test verbleibenden Stichprobenumfänge zur CML-Schätzung der Itemparameter sind ebenfalls Abbildung 23 (vorige Seite) zu entnehmen.

Erwartungswidriges Antwortverhalten ist besonders dann zu beobachten, wenn der Test -im Mittel- zu niedrige (oder zu hohe) Anforderungen an die Leistungsfähigkeit der Pbn stellt (s. Abbildung 24, nächste Seite).

Dieses motivationstheoretisch durchaus plausible Ergebnis ist nicht zuletzt ein gewichtiges Argument adaptiver, dem jeweiligen Fähigkeitsniveau der Pbn angepaßter Aufgabenauswahl. Die Analyse der Antwortvektoren zeigt dann auch, daß besonders große und schließlich zu globalem Pbn-Misfit führende Residuen z(vg) dann auftreten, wenn Pbn mit niedriger (und hoher) Fähigkeitsausprägung wohl zufällig sehr schwere Aufgaben richtig (bzw. eher leichte Aufgaben falsch) beantworten. In einer traditionellen Testentwicklung müßte dieses Ergebnis den Gültigkeitsbereich der Messungen auf eine "motivierte" Pbn-Population einschränken, gerade aber im Hinblick auf die vorgesehene adaptive Testung, die hohe Aufgabenbeantwortungsmotivation a priorisch wahrscheinlich macht, scheint diese Einschränkung nicht erforderlich.

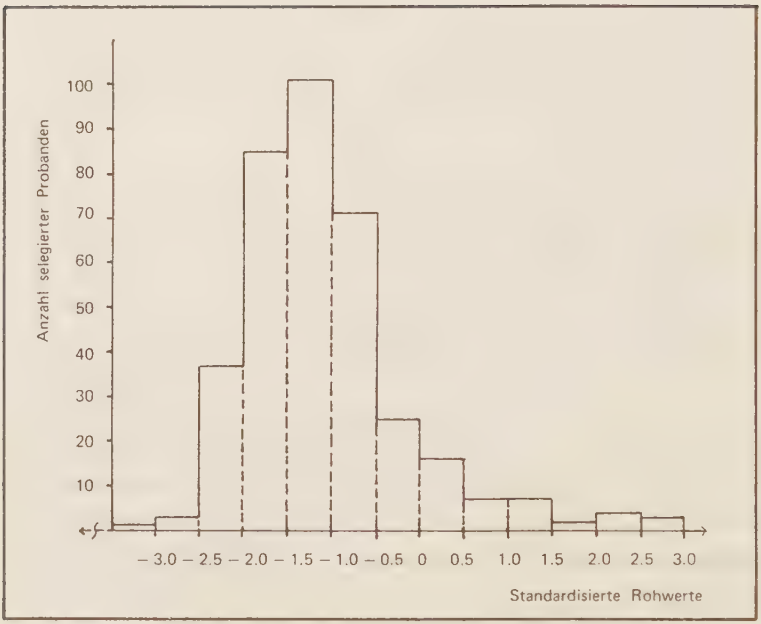


Abbildung 24 :
Verteilung der standardisierten Scores selegierter Probanden

10.1.2 DIE ZUSAMMENSTELLUNG HOMOGENER AUFGABENMENGEN

Die Itemmisfitanalyse mit der Residuenstatistik t(g) (vgl. Kp.5.3) steht bei unselegierten, d.h. nicht nach t(v) "homogenisierten" Probandenstichproben, ebenfalls vor dem Problem, die Quelle des Misfit zu lokalisieren. Einzelne, unerwartete Pbn-Reaktionen können einen Itemmisfit "vortäuschen". Dies sei an folgendem realen Fall verdeutlicht: In Test 7 weisen die beiden Items 14 und 24 mit den PROX-Schätzungen b(14)=2.44 und b(24)=1.67 hochsignifikante CHI**2- bzw. t(g)- Misfitstatistiken auf (s. Abb. 25).

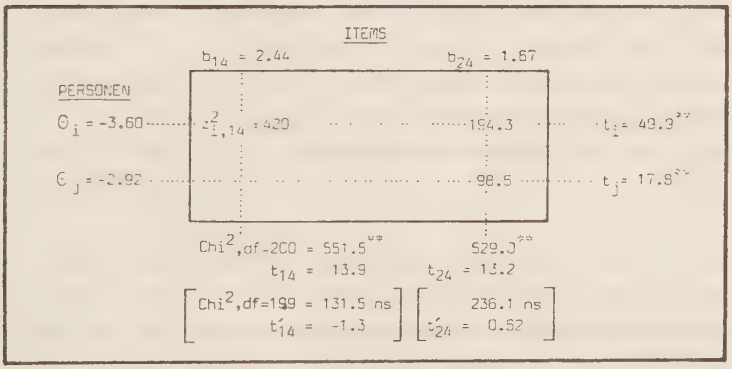


Abbildung 25:
Itemmisfit und zugehörige Statistiken

Die Analyse der Einzelbeiträge zu den CHI**2- bzw. z-Statistiken zeigt hier, daß allein durch die unerwartet richtigen Beantwortungen des "schweren" Items 14 durch die sonst meist falsch antwortenden Pbn mit $\Theta(1)$=-3.60 die Residuensumme für diese Aufgabe einen extrem hohen, signifikanten Wert annimmt. Pb i errät vermutlich auch die richtige Lösung des Items 24, was zwar einerseits zu einem signifikanten Pbn-Misfit (z(1)=249.9) führt, letztendlich aber auch das Item 24 über den Itemmisfit auffällig

werden läßt. "Bereinigt" man nun die Summenstatistiken um die durch die misfittenden Pbn i und j bedingten extremen Beiträge z**2, so resultieren nicht signifikante CHI**2- bzw. z-Statistiken (s. Abb. 25, eingeklammerte Werte) für diese Aufgaben.

Ein Plot der bivariaten Verteilung (Abb. 26, nächste Seite) aller 35x24=840 geschätzten Itemparameter und Anpassungsstatistiken t(g) zeigt, daß 81 Aufgaben zu selegieren wären, wollte man sich ausschließlich an dem kritischen Wert z=5 orientieren. Unter diesen hauptsächlich mittel- bis hochschwierigen Items befinden sich aber 13 Aufgaben, bei denen weniger als 5% aller Residuen (z(vg)) einen kritischen Wert von z(vg)**2=5 überschreiten. Wenige extrem hohe Einzelbeiträge, hauptsächlich durch wahrscheinlich zufällige Aufgabenlösungen von Pbn mit niedriger Fähigkeitsausprägung bedingt, lassen hier fälschlicherweise erhebliche Iteminhomogenitäten vermuten. Die nach BEIDEN Kriterien auffälligen 68 Aufgaben sind vermutlich nicht homogen im Sinne des RM.

Daß aus dem Pool der 616 verschiedenen Items nach dieser Modellkontrolle aber letztlich weniger als 68 Items zu selegieren sind, hat folgenden Grund: Die genannten Zahlen beziehen sich auf alle 840 Aufgaben der 35 Testversionen, d.h. gleiche, in zwei Tests vorkommende Linkitems wurden im Rahmen der Homogenitätskontrollen für die Einzelskalen zunächst wie verschiedene Items behandelt. Unter den 68 inhomogenen Items befinden sich 32 Tag- und 27 Linkitems. Von den 27 Links weisen 9 in beiden Tests signifikante Werte auf, so daß schließlich nicht 68, sondern 59 von 616 verschiedenen Items nach diesem Modelltest zu selegieren sind. Die übrigen 18 Linkitems fallen nur in jeweils einem der beiden Tests auf, nichtsdestoweniger sind sie aus dem Itempool auszuscheiden. Diese alternierenden Inhomogenitäten machen aber sehr deutlich, daß Modellkonformität und -inkonformität, so wie sie sich in diesem Anpassungstest darstellen, nicht ausschließlich und immer Qualitäten der Items per se wiederspiegeln, sondern mitunter auch

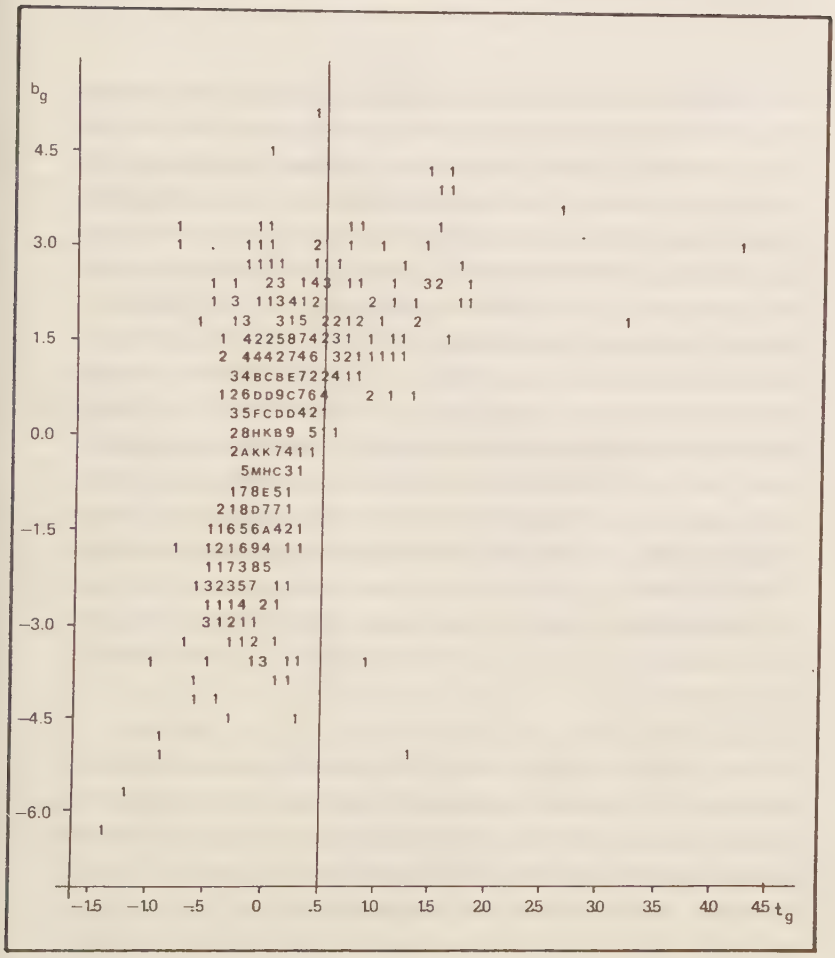


Abbildung 26 :
Bivariate Verteilung der PROX – Schwierigkeitsschätzungen b_g
und der Itemanpassungsstatistiken t_g

Schwierigkeitszusammensetzungen der Tests oder Stichprobenzusammensetzungen und Durchführungsbedingungen. Da dieser Problematik aber im vorliegenden Fall nicht nachgegangen werden kann, wird hier konservativ verfahren, d.h. auch "einfach" auffällige Linkitems werden ausgeschieden.

Die Konstanz der Itemparameterschätzungen ist besonders bezüglich, in der anvisierten Pobpulation schon augenscheinlich deutlich unterschiedlich ausgeprägter Probandencharakteristika bedeutsam. Die hier konstruierten Denkaufgaben sollen in der Auswahl- und Platzierungsdiagnostik zukünftiger Wehrpflichtiger und Freiwilliger der Bundeswehr eingesetzt werden. Diese Population weist weder Geschlechts- noch wesentliche Altersunterschiede, wohl aber beträchtliche Schulbildungsunterschiede auf. Die Bedeutungsinvarianz der Items gegenüber diesem Kriterium soll hier geprüft werden, indem für alle 35 Einzeltests die CML-Schätzungen der je 24 Itemparameter in der Gruppe der Wehrpflichtigen mit und ohne Hauptschulabschluß mit den entsprechenden Schätzungen in der Gruppe "Realschüler" und "Abiturienten" verglichen werden. Es wird davon ausgegangen, daß dieser Test auf "Schulbildungs-bias" der Matrizenitems trennscharf ist, da das Teilungskriterium vermutlich auch bezüglich der latenten Fähigkeitsdimension die Varianz zwischen den Teilgruppen heterogenisieren dürfte.

Auf explizite numerische und graphische Gegenüberstellung der aus insgesamt 70 Teilgruppen geschätzten 1.680 Schwierigkeitsparameter wird hier aus Gründen der Überschaubarkeit zugunsten numerischer Darstellungen der resultierenden Differenzenstatistiken u(g) (vgl. Kp. 5.3; Formel (32)) verzichtet.

Unter der Null-Hypothese wird behauptet, die Annahmen des RM seien gültig, was hier bedeutet, die Schwierigkeitsschätzungen gleicher Items seien bis auf Schätzfehler äquivalent.

Um einzelne überzufällig große Differenzen im Sinne dieser Modellkontrolle auf genuin schulbildungsspezifische Bedeutungsinstabilitäten der Item zurückführen zu können, soll auch hier von der nach 10.1.1 "homogenisierten" Probandenstichprobe ausgegangen werden.

Da für jedes Tagitem die Teststatistik z(g) und nach Partitionierungen der Stichprobe nach Schulbildung und Gruppen mit 'gleichen' Itembeantwortungen (s.u.) maximal 25 abhängige Differenzenstatistiken u(g) berechnet und im Hinblick auf eine Selektionsentscheidung interpretiert werden, wächst bei festgelegtem Alpha-Fehler das Risiko, irrtümlich ein Tagitem zu selegieren, beträchtlich. Dies Irrtumswahrscheinlichkeit ist für die Linkitems erheblich höher, da hier doppelt so viele, also maximal 52 abhängige Modellkontrollen durchgeführt werden. Würde der "Fehler erster Art" bei Alpha=.05/2 (zweiseitiger Test) festgelegt, bedeutete dies eine deutliche Verminderung des "Beta-Fehlers" und damit ein unnötig konservatives Vorgehen. Nach Vorschlägen von KRAUTH und LIENERT (1973) wurde hier das Signifikanzniveau von Alpha=.05 der Zahl abhängiger Tests angepaßt, so daß für die Bewertung der approximativ standardnormalverteilten u(g)-Statistiken kritische Werte von u(crit)=3.08 für die Tagitems und u(crit)=3.28 für Linkitems resultieren.

Bei insgesamt 24 Aufgaben mittlerer bis hoher Schwierigkeit weisen die Itemparameterschätzungen aus der "Abiturienten/ Realschülerstichprobe" und der "Hauptschülerstichprobe" erhebliche Unterschiede auf (s. Abb. 27, nächste Seite). Unter diesen Aufgaben befinden sich 14 Tagitems und 10 Linkitems, wobei hier wiederum auffällt, daß lediglich eine Linkaufgabe in beiden Tests erhebliche Schätzwertdifferenzen aufweist.

Zusätzliche Modelltests durch Teilgruppenbildung wie im vorliegenden Fall schienen schon deshalb gerechtfertigt, da sie offenbar gegenüber anderen oder zusätzlichen Verletzungen der Modell-

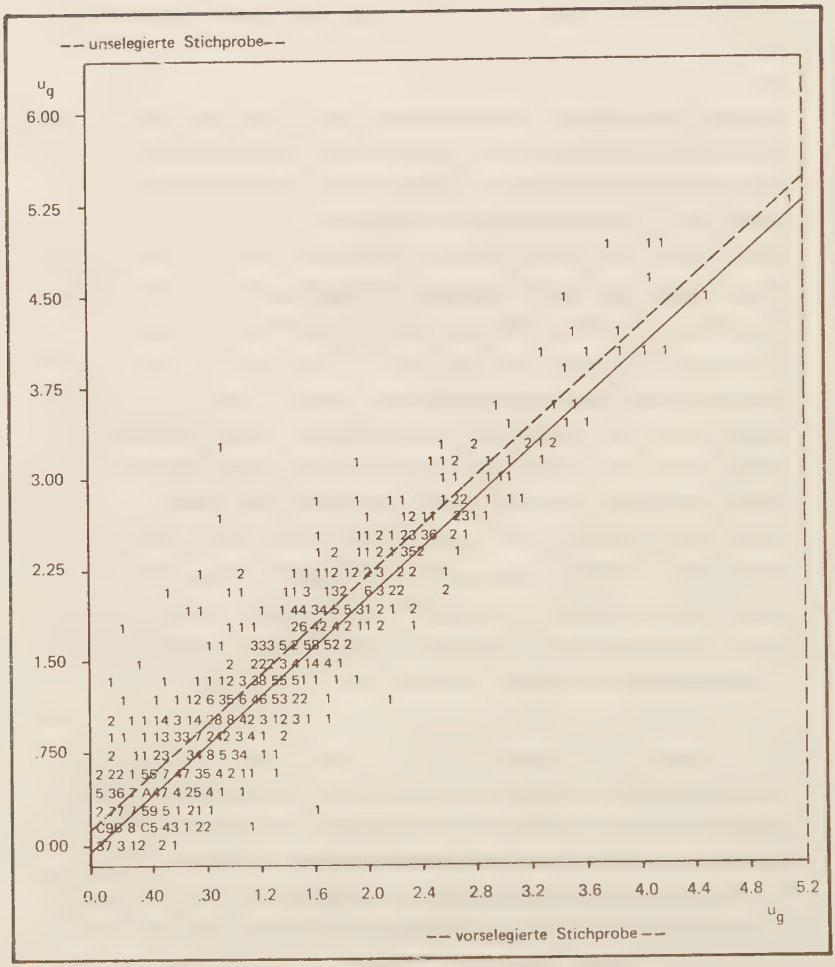


Abbildung 27 :
Verteilung der Differenzenstatistiken u_g aus den homogenisierten und ursprünglichen Probandenstichproben

annahmen sensitiver sind als die Teststatistiken z(g).

Abbildung 28 zeigt, daß lediglich 12 von 83 auffälligen Items in beiden Modellkontrollen als inhomogen ausgewiesen werden (III. Quadrant). Außerdem sind die u(g)- und t(g)-Statistiken "homogener" Items unkorreliert (II. Quadrant). Der Frage, inwieweit Beziehungen zwischen verschiedenen Arten "technischer" Selektionskriterien und strukturellen oder phänomenologischen Merkmalen selegierter Items bestehen, wird in Kapitel 11.3 nachgegangen.

Im folgenden soll der Vorschlag VAN DEN WOLLENBERG's (1979), Items als Teilungskriterien zu verwenden, in modifizierter Form aufgegriffen werden. Wegen des beschränkten Stichprobenumfangs (etwa n=200 pro Test) soll eine Teilgruppenbildung nach Items nur dann erfolgen, wenn die kleinere der beiden Gruppen mit gleichen Itembeantwortungen mit mindestens n=50 Pbn besetzt ist, d.h., daß bei einem Stichprobenumfang von n=200 nur nach Items mit relativen Lösungshäufigkeiten zwischen p=.25 und p=.75 partitioniert wird. Hier wird wieder deutlich, daß der Vorteil eines systematischen Vorgehens, d.h. hier die Teilung nach allen Items eines Tests, zugunsten von Praktikabilitätsüberlegungen aufgegeben werden muß. Die prinzipielle Motivation für dieses Vorgehen bleibt aber auch unter solchermaßen eingeschränkten Bedingungen erhalten, besonders dann, wenn Fitanalysen auf Itemebene durchgeführt werden. Nach der von FORMANN (1981) formulierten Logik ist die Submengenbildung nach Einzelitembeantwortungen noch am ehesten geeignet, Schätzwertinstabilitäten aufzudecken. Sollen Schätzwertdifferenzen in diesem Kontext mittels der u-Statistiken von FISCHER und SCHEIBLECHNER (1980) bewertet werden, so ist zu beachten, daß für jedes Item g aus überlappenden Stichproben so viele Statistiken z(g) berechnet werden, wie Kriterienitems im Range .25<p<.75 vorhanden sind.

Für die multiplen Partitionierungen wurden insgesamt 386 "Kriterien" - Items mit relativen Lösungshäufigkeiten .25<p<.75 heran-

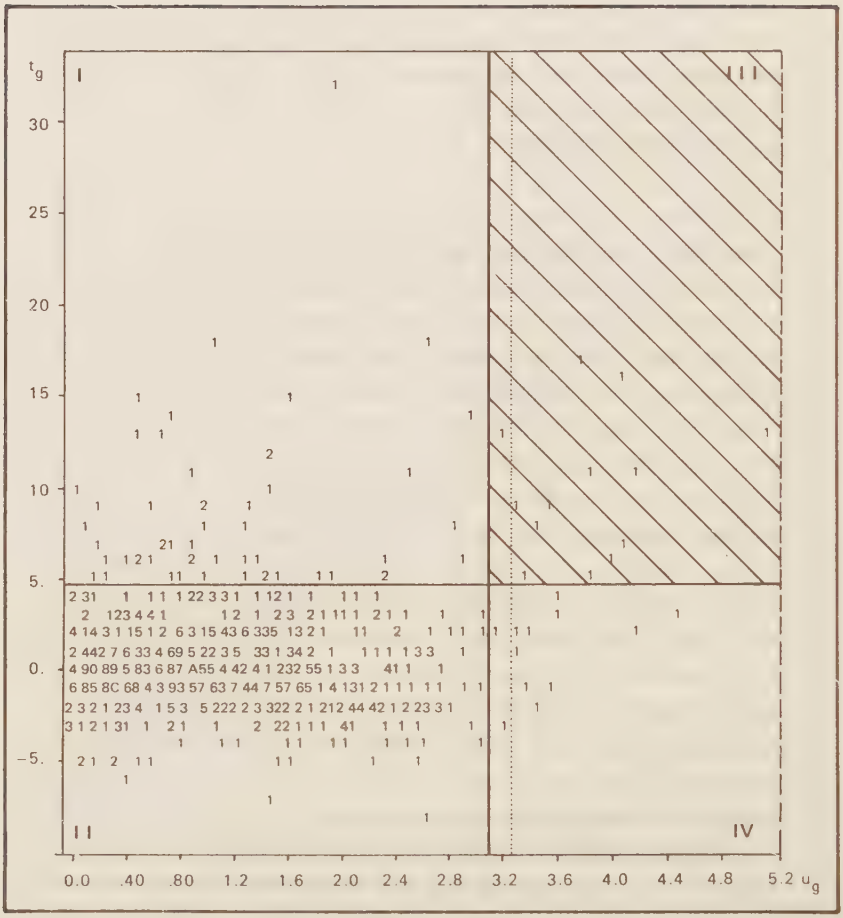


Abbildung 28 :
Gegenüberstellung der geschätzten u_g – und t_g – Statistiken

gezogen. Pro Test konnten entsprechend der Zahl geeigneter Kriterienaufgaben zwischen 5 und 16 (im Mittel 11) Partitionierungen vorgenommen und entsprechend zwischen 115 und 368 Anpassungsstatistiken u(g) für jedes Item berechnet werden (vgl. Tabelle 4, nächste Seite).

Von 8.878 Differenzenstatistiken wiesen 69 signifikante Werte auf (u(g;tag)>3.08 und u(g;link)>3.28), wobei zu beachten ist, daß 15 Items bei mehr als einer Partitionierung und 42 Aufgaben nur jeweils einmal auffällig sind. Die 35 Tag- und 22 Linkitems mit instabilen Schätzwerten verteilen sich ungleichmäßig auf die 35 Tests. Wie Tabelle 4 zu entnehmen ist, wären allein nach diesem Modelltest zwischen 0 und 7 Items pro Test zu selegieren, wobei die Anzahl auffälliger Items in einem Test nicht mit der Zahl statistischer Kontrollen korreliert ist.

66% dieser Aufgaben werden ausschließlich durch die multiplen Stichprobenteilungen nach Kriterienitems als inhomogen ausgewiesen, 19 Items erwiesen sich bereits in anderen Kontrollen als nicht modellkonform. Es ist allerdings nicht auszuschließen und eher wahrscheinlich, daß ein -allerdings hier nicht möglichessystematisches Vorgehen, das alle Aufgaben eines Tests als Teilungskriterien einbezöge, "trennschärfer" wäre und weitere Inhomogenitäten aufgedeckt hätte. Außerdem ist zu bedenken, daß die Teststärke ("power") des verwendeten Tests bei "leichten" und "schweren" Kriterienitems geringer ist als bei mittelschweren Aufgaben. Die in diese Prüfgröße u(g) (vgl. Kp. 5.3: Formel 32) eingehenden Varianzen der Schwierigkeitsparameterschätzungen sind abhängig von der Adaptivität der Testaufgaben und dem Stichprobenumfang. Je größer die Teilstichprobe und je mehr die Schwierigkeit der Aufgabe g den Fähigkeitsausprägungen der Pbn angemessen ist, desto kleiner werden die Schätzwertvarianzen und desto schärfer ist der Anpassungstest. Es ist also nicht auszuschließen, daß unter den gegebenen Randbedingungen selbst für praktische Belange intolerable Schätzwertinstabilitäten allein

Test	Anzahl der Kriterien-Items	Anzahl seleg. Items
1	11	1
2	7	4
3	8	1
4	11	2
5	15	0
7	15	2
8	10	1
9	9	3
10	9	2
11	11	1
12	14	0
13	11	0
14	11	2
15	5	1
16	11	2
17	12	0
18	12	2
19	11	1
20	13	0
21	9	3
22	7	0
23	13	1
24	14	2
25	16	2
26	16	2
27	6	5
28	11	2
29	11	0
30	12	1
31	16	0
32	6	6

Fortsetzung Tabelle 4:

Test	Anzahl der Kriterien-Items	Anzahl seleg. Items
33	7	7
34	16	0
35	12	0

Tabelle 4:
Anzahl selegierter Items pro Test

durch die multiplen Stichprobenteilungen nicht entdeckt werden konnten. Folgende Zusammenstellung zeigt aber, daß das Verfahren gerade in einem zusammenhängenden Geflecht verschiedener Modelltests einen bedeutsamen und eigenständigen Beitrag zur Zusammenstellung homogener Skalen leistet.

'Technische' Selektionskriterien	Anzahl selekt. Items	Anteil an Gesamtzahl	
RES + SCHULB + MP :	10	9.2	
RES + SCHULB :	3	2.8	19.3
RES + MP :	3	2.8	
RES :	44	40.4	
SCHULB :	6	5.5	80.7
MP :	38	34.8	
TOTAL :	109	100.0	

Tabelle 5:

Verteilung der selegierten Items nach verschiedenen 'technischen' Kriterien. Hierbei bedeuten:

RES : Anpassungstest mittels der Residuenstatistik

SCHULB: Modelltest mittels Teilung der Stichproben nach dem externen Kriterium Schulbildung

MP : Modelltest mittels Teilung der Stichproben nach Gruppen gleicher Itembeantwortungen

Von 616 verschiedenen Matrizenaufgaben sind nach allen Homogenitätskontrollen für die 35 Einzelskalen 109 Items (=17.7%) zu selegieren. Über 80% inhomogener Aufgaben sind jeweils nur in einem der Modelltests auffällig. Dabei wurden allein durch das Verfahren 'MP' immerhin 34.8% dieser Items als inhomogen identifiziert, was diesen zusätzlichen aufwendigen Einsatz im nachhinein legitimieren mag. Daß die durchgeführten Kontrollen unterschiedliche Aspekte der Modellkonformität thematisieren, wird schon dadurch deutlich, daß lediglich 10 Items in allen und 11 gleichzeitig in jeweils zwei Modelltests auffällig sind. Sind die in Kapitel 5.3 diskutierten theoretisch befriedigenden Modellkon-

trollen, wie im vorliegenden Fall wegen zu kleiner Stichprobenumfänge nicht durchführbar, empfiehlt sich offensichtlich ein mehrstufiges Vorgehen, das Kontrollen mit unterschiedlichen Ausgangspunkten und Zielsetzungen einschließt.

Die bisher durchgeführten Anpassungstests sollen in erster Linie zur Beantwortung der Frage führen, ob und welche Aufgaben der 35 Tests homogene Einzelskalen im Sinne des RM konstituieren. In einem weiterführenden Analyseschritt ist nun zu prüfen, ob die Homogenitätshypothese für den verbleibenden Matrizenaufgabenpool beibehalten werden kann, d.h. ob alle verbleibenden 507 strukturverschiedenen Aufgaben eine einzige latente Fähigkeitsdimension aufspannen.

11.1.3 LINKFIT-ANALYSE UND ÜBERBLICK ÜBER DIE KONTROLLEN ZUM RM

Im vorliegenden Design sind jeweils zwei "benachbarte" Tests durch jeweils 6 Aufgaben in einer Linkstruktur miteinander verkettet (vgl. dazu die Ausführungen in Kp. 9). Eine wesentliche Voraussetzung zur Berechnung von Translationskonstanten, die die Skalen ineinander überführen ("linkage"), ist die Homogenität der Linkitems. Nach den bisherigen Modelltests verbleiben mindestens zwei, im Mittel vier und maximal alle sechs Linkitems für jede Testverkettung. Testübergreifender Modellfit soll hier nach Vorschlägen von WRIGHT und STONE (1979) überprüft werden, indem die unabhängigen Schätzungen der Linkitems aus jeweils zwei Datensätzen miteinander verglichen werden. Sind zwei Tests auf einer gemeinsamen Skala kalibrierbar, so sind die Differenzen zwischen den unabhängigen Schätzungen der Linkitems bis auf Schätzfehler gleich groß. Der Misfit eines einzelnen Linkitems drückt sich daher in überzufällig großen (standardisierten) Abweichungen z(RD) von der mittleren Differenz zwischen den Schätzwerten (vgl. Formel (44) aus.

Bilden die Items einen brauchbaren Link, so sind diese Residuen standardnormalverteilt mit N(0,1) bei $K \rightarrow \infty$. Die Qualität bzw. Anpassung des gesamten Link zweier Tests kann somit bei einer ausreichenden Anzahl von Linkitems über den Erwartungswert und die Standardabweichung der z(RD)-Verteilung abgeschätzt werden. Nachdem bereits 50 Linkitems nach Anpassungskontrollen der Einzeltests zu selegieren waren, sind nunmehr aus 7 Links (s. Markierung (V) in Abb. 29 auf der nächsten Seite) weitere 11 Aufgaben auszuschließen, da bei diesen Items erheblich größere Schätzwertdifferenzen auftreten, als dies bei Berücksichtigung der Schätzfehler zu erwarten ist (z(RD)>2.56 bei Alpha=.01). Dies hat zur Folge, daß drei Links (13/14, 31/32, 32/33) aus nur zwei bzw. drei Aufgaben bestehen, was die Angabe und Interpretation von "Verteilungs-Charakteristika" der Residuen natürlich nicht sinnvoll erscheinen läßt. Um diese Skalen anzugleichen, können

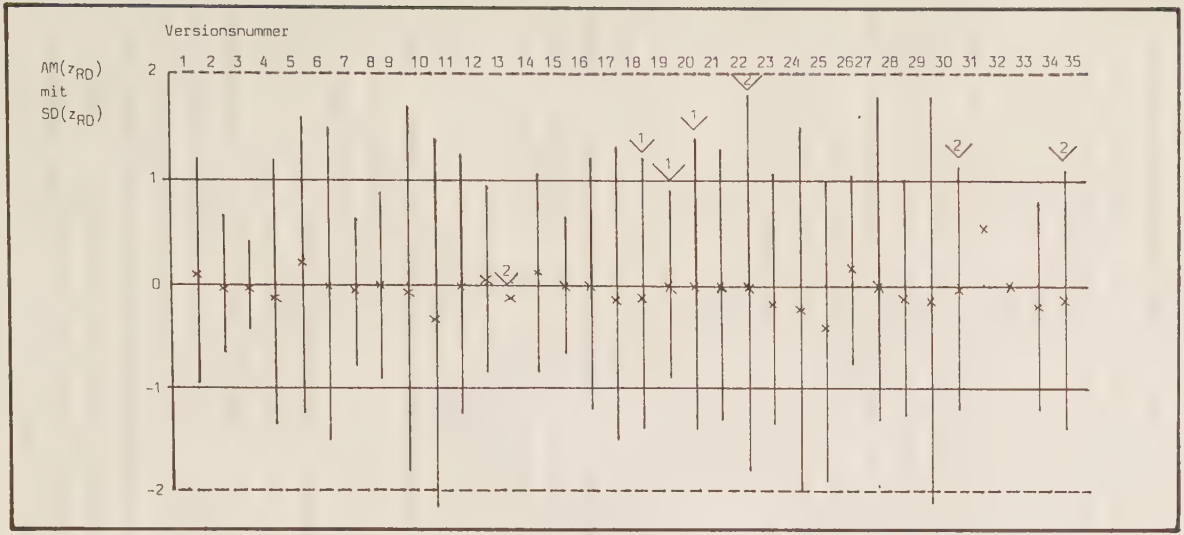


Abbildung 29 :
Arithmetisches Mittel (AM (z_{RD})) der Residualwerte z_{RD}
und deren Standardaweichungen (SD (z_{RD})) bei 35 Links

dennoch aus den wenigen verbleibenden Linkitems Transformationskonstanten berechnet werden; ihre Reliabilität ist jedoch unbestimmt. Die für die übrigen 27 Links berechneten Kennwerte der Residuenverteilungen (vgl. Abb. 29) sind zumindest tendentiell interpretierbar und lassen eine hinreichende Gültigkeit der Links vermuten.

Abbildung 30 (s. nächste Seite) zeigt, wie sich die insgesamt 134 "problematischen" Aufgaben auf die 35 Testversionen und die jeweils 24 Itempositionen verteilen. Die Zusammenstellung berücksichtigt bereits weitere zwei Link- und sieben Tagitems, für die, wegen zu geringer relativer Lösungshäufigkeiten $p(g)<.03$, keine zuverlässigen CML-Endschätzungen, durchgeführt nach Selektion aller "inhomogenen" Items (und Probanden), gewonnen werden konnten.

Geschwärzte doppelt breite Kästchen bezeichnen Linkitems, die in beiden Tests auffällig wurden. Etwa die Hälfte der breiten Kästchen ist halb geschwärzt und kennzeichnet alternierend inhomogene Links. Es wird unmittelbar deutlich, daß bei bestimmten Tests, wie auch auf bestimmten Positionen in der Aufgabenreihe deutlich mehr inhomogene Matrizenaufgaben zu finden sind als anderswo.

Wie der unteren Randverteilung zu entnehmen ist, werden in den 35 Tests zwischen 0 und 10 Aufgaben entweder in Anpassungskontrollen oder in der Linkfit-Analyse als nicht modellkonform ausgewiesen. Die stark variierende "Testkonformität" geht einher mit der beobachteten Variabilität des "Personenstichprobenfit". Da alle Tests weitgehend strukturparallel und die Stichprobenzusammensetzungen ähnlich sind, liegt die Vermutung nahe, daß an einigen Erhebungsorten die notwendigen Voraussetzungen für vergleichbare und der vorliegenden Problemstellung angemessene Testdurchführungen nicht hergestellt wurden. Sie wird gestützt durch einige der Angaben, die sich in vorliegenden Erfahrungsberichten aus 21 Kreiswehrersatzämtern (KWEA) finden. Es fällt beispielsweise auf,

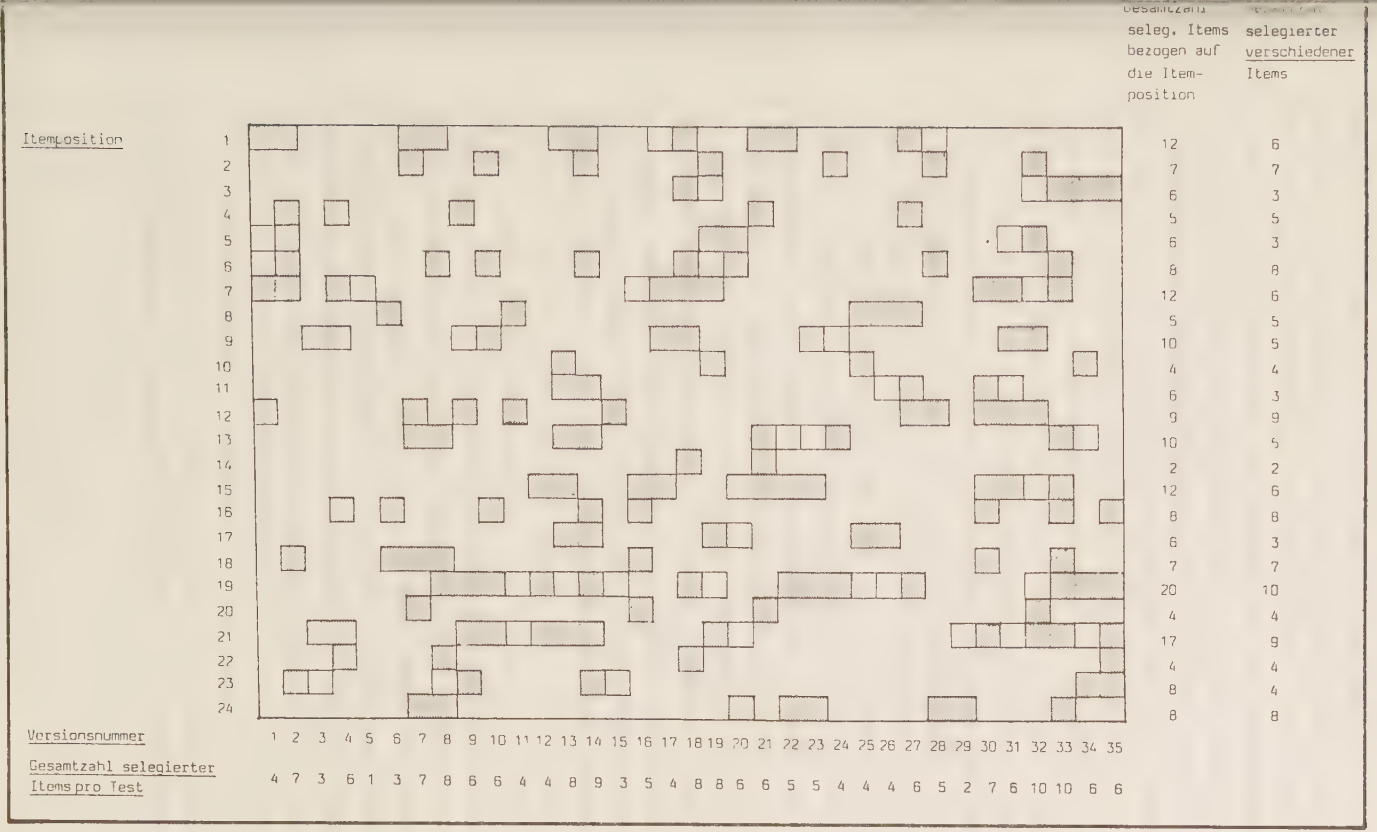


Abbildung 30 :
Verteilung der selegierten Items auf Testversionen
und Itempositionen

daß die aus dem KWEA München berichtete durchschnittliche Testbearbeitungszeit für die Matrizentestversion 33 mit 35 Minuten (Median: 20 (!) Minuten) ganz erheblich unter den übrigen berichteten Durchschnittszeiten liegt, die sich um 45 bis 50 (!) Minuten bewegen. Version 33 zeigt aber auch den ungünstigsten Modellfit aller 35 Tests, da insgesamt 10 Items allein aus dieser Version zu selegieren sind. 7 der 10 auffälligen Items gehören der zweiten Testhälfte an, die erste Testhälfte weist hingegen unterdurchschnittlich wenige Inhomogenitäten (3) auf. Ähnliches ist im übrigen auch bei den ebenfalls stärker auffälligen Versionen 7, 8 und 14 zu beobachten. Dies läßt darauf schließen, daß an den entsprechenden Erhebungsorten und -zeitpunkten, entgegen der Testleiteranweisung, nicht allen Pbn ausreichende Bearbeitungszeiten eingeräumt wurden. Speed-Einflüsse in der zweiten Testhälfte dürften hier, besonders aber bei den letzten Items wirksam werden und somit die Leistungsgeschwindigkeit der Pbn in zunehmendem Maße den Lösungserfolg bei diesen Aufgaben mitbestimmen. Da aber das logistische Meßmodell von RASCH die Wahrscheinlichkeit einer Antwort als Funktion der eindimensionalen latenten Fähigkeit auffaßt, werden bei Interaktionen dieser Fähigkeitsdimensionen mit der Leistungsgeschwindigkeit der Pbn zentrale Modellannahmen verletzt. Wie gezeigt werden konnte, sind die hier durchgeführten Gültigkeitskontrollen zum RM offenbar sensitiv gegenüber dieser Art von Modellverletzungen.

Die 134 selegierten Aufgaben verteilen sich auch recht ungleichmäßig auf die 24 Itempositionen. Die vermutlichen Speed-Einflüsse bei einigen Tests spiegeln sich zumindest tendentiell auch in der Verteilung selegierter Items auf die Itemposition. Aus den letzten Testdritteln mußten mehr Aufgaben (k=49) selegiert werden als aus den ersten Dritteln (k=43) und den mittleren Dritteln (k=42).

Interessanterweise stehen die Gipfel der rechten Randverteilung im Zusammenhang mit strukturellen Merkmalen der Aufgaben, die ja

gemäß einem Strukturverteilungsschlüssel (Abb. 20 und 21 in Kp. 9) auf die entsprechenden Positionen verteilt wurden. Auf den Positionen 6, 12 und 24 finden sich beispielsweise immer Tagitems, die mit "integrierten Repräsentanten" (IR) realisiert wurden und auf den Positionen 19 und 21 Linkitems, von denen 2/3 IR und 1/3 "integrierte Komponenten" (IC) enthalten. Diese Beobachtung gibt erste, noch recht unspezifische Hinweise auf kritische Elemente der zugrundeliegenden Konstruktionstheorie. Sie ist nunmehr Anlaß zu einer systematischen Analyse struktureller Merkmale problematischer Aufgaben (s. Kp. 11.3).

11.2 DIE ZUSAMMENSTELLUNG DER COMMONSCALE

Die Angleichung der 35 Einzelskalen stellt den vorläufig letzten Teilschritt in der Entwicklung einer homogenen "Breitband-Skala" (vgl. LORD, 1977) dar. Aus den verbliebenen Linkitems wurden die 34 Translationskonstanten nach Formel 44 erneut geschätzt. Wie nachfolgende Abbildung zeigt, weisen die Shifts je nach Schwierigkeitszusammensetzung der Einzeltest Beträge zwischen -0.32 und 1.88 auf.

Aus den Tests 23 bis 31 wurden zumeist weniger schwere Aufgaben selegiert als aus den anderen Testversionen. Die nach den CML-Endschätzungen der Itemparameter durchgeführten Produktnormierungen führen nun gerade bei diesen Skalen dazu, daß die Schwierigkeitsschätzungen betreffender Linkitems vergleichsweise -d.h. bei testübergreifender Betrachtung- geringere Beträge aufweisen. Die Differenzen zwischen jeweils bereits angeglichenen Linkitemschwierigkeiten des einen Tests und

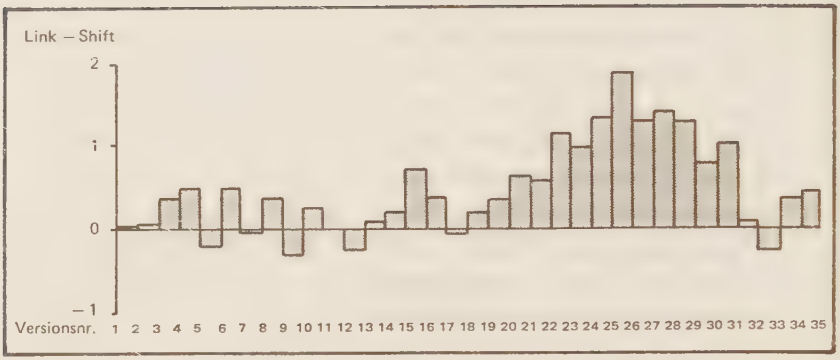


Abbildung 31 :
Die 34 Translationskonstanten im Überblick

den entsprechenden Schätzwerten des anzugleichenden Tests nehmen daher, wie ihr arithmetisches Mittel positive Beträge an.

Die Angleichung der 35 Einzelskalen erfolgte nun, indem, beginnend bei Testversion 1 und 2, dann sukzessive um je einen Test fortschreitend die Link-Shiftkonstanten zu den Schwierigkeitsschätzungen des anzugleichenden "nächsten" Tests addiert wurden. Nach jedem Shift wurden die endgültigen Lokalisationen der Linkitems auf der Schwierigkeitsskala durch Mittelung der beiden jeweils verfügbaren Schätzer bestimmt.

Die 502 Itemschwierigkeitsschätzungen (b(Pool)) sind auf der zentrierten gemeinsamen Skala ("Commonscale") zwischen b(g)=-4.05 und b(g)=4.28 annähernd normalverteilt mit AM(b(g))=0 und der Standardabweichung SD=1.45. Die Schätzwerte "leichter", besonders aber extrem "schwerer" Items sind -erwartungsgemäß- mit größerer statistischer Unsicherheit behaftet als "mittelschwere" Items. Fast 90% aller Itemparameter konnten mit praktisch hinreichender Genauigkeit SD(b(g))<0.25 geschätzt werden. Gerade die Schwierig-

keitsschätzungen hochkomplexer Aufgaben bedürfen aber weiterer Absicherung; nicht zuletzt um die statistische "Power" itembezogener Anpassungstests zu erhöhen und so die Gültigkeitsannahme für die Items zu untermauern.

11.3 DIE PROBLEMATIK INHOMOGENER ITEMS

Obwohl die vorläufig angenommene meßtheoretische Brauchbarkeit von über 550 Matrizenaufgaben zweifellos schon jetzt für einen an objektiven Regeln orientierten Konstruktionsansatz sprechen, scheint er jedoch in seiner Ausgangsformulierung nicht uneingeschränkt gültig zu sein. Spezifikationen, Revisionen oder Erweiterungen setzen die Identifiaktion einer verallgemeinerbaren und in (Konstruktions-) Regelbegriffen formulierten inhaltlichen Systematik voraus, mit der die oben und in weiteren Anpassungskontrollen technisch selegierten Aufgaben eindeutig kategorisierbar sind. Hier soll zunächst davon ausgegangen werden, daß sowohl konstruktionstheoretisch explizierte, wie auch vordefinitorische Aufgabenmerkmale, Störfaktoren darstellen können, die letztlich Inhomogenitäten bedingen. Da bereits oben nicht-aufgabenimmanente Inhomogenitätsquellen (z.B. Speed-Einflüsse) aufgezeigt wurden, ist darüberhinaus davon auszugehen, daß für betroffene Items strukturorientierte Erklärungen zwangsläufig scheitern. Das bestehende Rational wäre gezielt und einfach revidierbar, ließe sich die mangelnde Konformität der ausgeschiedene Aufgaben eindeutig auf bestimmte objektivierte logische Itemstrukturen zurückführen. Konstituierende Elemente des Konstruktionsansatzes sind die "Realisierungsmodi", die "geltenden Regeln" und die "Richtungen der Regelgeltung". Konstruktionsregeln bzw. Operationen, die zu Inhomogenitäten führen, sollten überproportional häufig unter den 134 selegierten Aufgaben zu finden sein.

Abbildung 32:
Absolute und relative Häufigkeiten (geschwärzt)
der 14 Operationen bei den selegierten Items

Betrachtet man die auf ihre Gesamthäufigkeit im Itempool relativierten Operationshäufigkeiten in selegierten Items (vgl. geschwärzte Säulen in Abbildung 32 auf der vorigen Seite), so sind kaum einzelne besonders problematische Regeln/ Operationen zu erkennen.

Tendentiell werden aber mehr Items auffällig, die mit "integrierten Repräsentanten" (IR) realisiert werden und solche, bei denen die "Variation offener Gestalten" (Vo) zur richtigen Lösung führt. Unter den 134 selegierten Aufgaben befinden sich 39, in denen diese Regeln/ Operationen enthalten sind und in allen 39 Items treten sie in Kombination miteinander auf. Da dies nun bei weitem nicht alle Aufgaben des Pools sind, die diese Regelkombination realisieren, muß es sich hierbei um Konkretionen handeln, die sich von den übrigen "homogenen" IR-Vo-Items in zu bestimmender Hinsicht unterscheiden.

Abbildung 33 zeigt eine der inhomogenen IR-Vo-Matrizen (A) und ein entsprechendes homogenes Item (B).

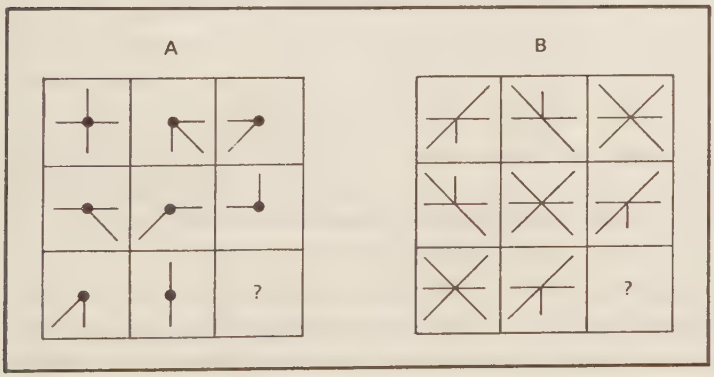


Abbildung 33 :
Inhomogenes IR – Vo – Item (A) und entsprechendes homogenes Item (B)

Im Beispiel A sind eine "Subtraktion" (Su) und eine "Variation offener Gestalten" (Vo) mit "integrierten Repräsentanten" (IR) realisiert. Die Repräsentanten beider Regeln (Su und Vo) bilden bei diesem Realisierungsmodus eine "Gestalt", die von den Pbn durch eine "Suchoperation" aufzulösen ist. Das problematische Item A unterscheidet sich von den homogenen Aufgaben B dadurch, daß die Teilfiguren, auf die jeweils eine der beiden Regeln/ Operationen anzuwenden ist, ausschließlich über eben diese Regeln identifizierbar sind. Definitionsgemäß konstituiert die geltende Regel hier das einzige erkennbare Ordnungsprinzip für die Teilfiguren. Würden die beiden konstitutiven Komponenten des Items A in modifizierter Form separat repräsentiert (SC), entfiele die vorbereitende Gestaltauflösungsoperation als Teil des rationalen Aufgabenlösungsprozesses.

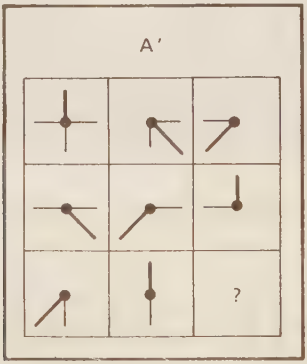


Abbildung 34 :
Beispielaufgebe A aus Abbildung 33
mit separierten Komponenten

Um diese Modifikation zu verdeutlichen wurden in obiger Darstellung A' die für die Regel-/ Operationskombination Vo-waagrecht/senkrecht relevanten Figurteile durch breitere Striche aus der Gesamtfigur "herausgelöst". Durch den so veränderten Realisierungsmodus (IR→SC) wird die Aufgaben-/ Problemstruktur deutlich verändert, da hier bereits phänomenologisch ein Hinweis auf

regelrelevante Repräsentanten erfolgt.

Die homogene "Variante" integrierter Repräsentation (vgl. Beispielaufgabe B) erfordert dagegen nach wie vor die intendierte Suchoperation zur Gestaltauflösung, da keine redundanzerzeugenden graphischen Hinweise auf relevante Merkmale existieren. Hier sind im Gegensatz zu Aufgaben des Typs A neben den jeweils geltenden Regeln spezifische Anordnungsmerkmale für Teilfiguren benennbar. So sind in Item B, das zwei waagerecht und senkrecht geltende Vo in einer "geschlossenen Gestalt" (IC) realisiert, die zusammengehörenden variierenden Repräsentanten in jeder der 8 Figuren diagonal angeordnet. Ordnungskonzepte wie "Diagonalität", "oben", "unten", "links", "rechts" und ähnliches sind kognitiv repräsentiert und können den Suchprozeß zur Gestaltauflösung anleiten. Im Lösungsprozeß von Aufgaben des Typs B kann der so initiierte Suchprozeß erfolgreich sein, in den heterogen Typ-A-Items hingegen nicht. Da sich die homogenen IR-Vo-Aufgaben von entsprechenden heterogenen Items -auf erkennbare Weise- nur durch solche impliziten zusätzlichen Ordnungsstrukturen unterscheiden, kann gefolgert werden, daß Matrizenitems wohl nur dann homogen sind, wenn die, in ggfls. erforderlichen "Suchoperationen" aktualisierten Ordnungsschemata auch Ordnungsmerkmale relevanter Figurteile darstellen.

Die hiermit verbundene inhaltliche Spezifikation des gemessenen Merkmals ist nicht nur im Rahmen der vorliegenden Itempoolkonstruktion von Bedeutung. Sie erklärt auch -zumindest ansatzweise- einen Teil der "technischen" Itemselektion in den Vorgängeruntersuchungen FORMANN's (1973) und PISWANGER's (1975) und konkretisiert zugleich die weniger auf meßtheoretische Aspekte bezogenen denkpsychologischen Untersuchungsergebnisse PUTZ-OSTERLOH's (1981, S.166ff.), die gerade in diesem Zusammenhang auf "differierende Problemlösungsschritte" (S.170) bei Matrizenitems aufmerksam macht.

In weiteren 35 Aufgaben konnten eindeutige Konstruktionsfehler identifiziert werden (doppelte oder fehlende Richtiglösungen, zwei gleiche Distraktoren oder ähnliches), obwohl die Items gewissenhaft nach den Vorgaben des Konstruktionsrationals erstellt und anschließend überprüft wurden. Bedenkt man allerdings, daß in kurzer Zeit von jedem Mitarbeiter über 100 Items entworfen, korrigiert und graphisch exakt dargestellt werden mußten, erscheint die Fehlerquote von ca. 5% akzeptabel.

Schließlich wurden aber "vorsorglich" 29 weitere Items selegiert, die sich zwar als homogen im Sinne des RM erwiesen hatten, andererseits aber aus verschiedenen Gründen nicht exakt den Konstruktionsregeln entsprachen. Es handelt sich hierbei NICHT um Konstruktionsfehler in o.a. Sinne, sondern eher um möglicherweise problematische "Auslegungen" der Konstruktionsvorschriften durch die Itemkonstrukteure.

Definitionsgemäß wird bei "Subtraktionen" von einer vollständigen Figur in der ersten ODER zweiten Spaltenposition eine Teilfigur ausgeblendet.

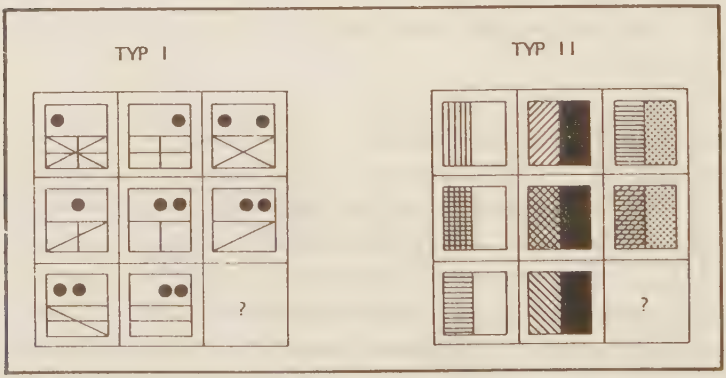


Abbildung 35 :
Beispiele für TYP I – und TYP II – Subtraktionen

Bei der Konstruktion wurden daher definitionsgemäß Items realisiert, bei denen sich die vollständige Figur entweder in der ersten (TYP I, in Abb. 35 linkes Item, untere Figurhälften, RR:w/s) oder in der zweiten Position befindet (TYP II, rechtes Item, linke Figurhälfte, RR:s). In einigen Fällen treten beide Realisierungsformen (TYP I und II) innerhalb EINES Items auf, was allerdings eindeutig der Konstruktionsvorschrift widerspricht. Diese Aufgaben erwiesen sich denn auch in der Regel als nicht modellkonform und wurden der Konstruktionsfehlerkategorie zugeordnet. Um die Interpretierbarkeit dieser Regeln zu verbessern und mögliche Uneindeutigkeiten aufzulösen, wurden in nachfolgende LLTM-Analysen nur TYP I-Subtraktionen einbezogen, bei denen sich die vollständige Figur in der ersten Zeilen- und/oder Spaltenposition befindet. 27 TYP II-Subtraktionen blieben unberücksichtigt, da sie über 35 Testversionen verteilt, gegenüber allen anderen Regeln/ Operationen deutlich unterrepräsentiert sind.

An dieser Stelle sei auf eine interessante Parallele aus dem Bereich arithmetischer Aufgabenlösungsfähigkeiten hingewiesen. BALLA und MACDONALD (1985) untersuchten die Dimensionalität eines "Arithmetic Skill Test", der verschiedenste Aufgaben zu allen vier Grundrechenarten enthält, indem sie ebenfalls die Fit-Statistiken von WRIGHT und STONE (1979) verwendeten. Dabei erwiesen sich Items mit bestimmten, definierten Facetten (vgl. GUTTMAN und SCHLESINGER, 1967) als inhomogen:

> "1. 2-digit minus 2-digit problems expressed in vertical format requiring borrowing;
> 2. Multiplication expressed in vertical format;
> 3. Multiplication by zero;
> 4. SUBTRACTION IN HORIZONTAL FORMAT WITH UNKNOWN IN POSITION A; and
> 5. Multiplication involving multipliers from 6 to 10" (BALLA und MACDONALD, 1985, S. 195; Hervorhebung von MH).

Der unter Punkt 4 genannte Typus von arithmetischen Subtraktionsaufgaben weist ohne Zweifel Ähnlichkeiten mit den hier problematisierten figuralen "TYP II-Subtraktionen" auf. Die naheliegende Hypothese, daß den Lösungsprozessen HOMOGENER Aufgaben beider Gegenstandsbereiche eine gemeinsame latente Fähigkeit zugrundeliegt, wäre in einer Untersuchung zu prüfen, in der beide Aufgabentypen EINER Personenstichprobe vorgelegt werden. Würde es dabei gelingen, die Lösungsprozesse auf eine gemeinsame Operationsstruktur zurückzuführen, so wäre dies zweifellos ein eindrucksvoller Beleg für die theoretische und praktische Bedeutsamkeit dieser Konstruktrepräsentation.

Bei einem kleinen Teil der selegierten Aufgaben (5) führt eine regelkonforme Realisation des Darstellungsmodus "integrierte Komponenten" (IC) zu einer inhomogenitätserzeugenden Verfälschung der jeweils "geltenden Regeln" (IC-Fehler).

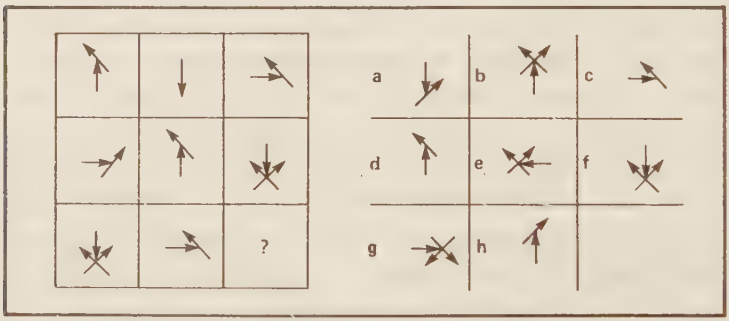


Abbildung 36 :
Beispiel für ein IC – Fehler – Item

Bei obigem Item werden im Zentrum der Einzelfiguren "offene Gestalten" (Pfeile) waagerecht/senkrecht variiert. Die Pfeilspitze ist hier ein relevantes Repräsentantenmerkmal. Die zweite Komponente besteht aus jeweils an den Pfeilspitzen der ersten

Komponente angeordneten, waagerecht/senkrecht "addierten" Pfeilen. Ihre Raumlage -als irrelevant gedachtes Repräsentantenmerkmal- ist dabei gemäß IC-Definition von einem relevanten Merkmal der ersten Komponente abhängig (vgl. die Definieten von IC auf S. 136). Die variable räumliche Anordnung der zu addierenden Figurelemente erschwert aber offenbar die notwendige Additionsoperation zumindest für einen Teil der Pbn. Durch die so ausgeführte Komponentenintegration entsteht ein Repräsentant einer Komponente (...eben nicht nur ...) dadurch, daß die beiden restlichen, in derselben Zeile/Spalte angeordneten Repräsentanten übereinandergelegt gedacht werden ("Addition", s. S. 128). Die veränderte Raumlage der "offenen Gestalten" bildet gewissermaßen einen "Störfaktor", der die Operation " Addition" durch zusätzlich erforderliche räumliche Orientierungsleistungen komplexer erscheinen läßt, als dies üblicherweise bei diesem Operationstyp der Fall ist.

Werden in integrierten Komponenten (IC) Regeln mit "offenen Gestalten" realisiert, so sollte deren Positionierung in der Gesamtfigur unverändert beibehalten werden, es sei denn, die Raumlageveränderung ist konstitutiv für die geltende Regel selbst (Seriation).

Bei 14 selegierten, extrem "leichten" Matrizenaufgaben, für die auch z.T. wegen zu geringen relativen Lösungshäufigkeiten keine Schwierigkeitsparameter geschätzt werden konnten, führen alternative, nicht-analytische Strategien zur richtigen Lösung. Es handelt sich hierbei, wie in Abbildung 37, um einige der Items, in denen die Regel "Identität" (I), "waagerecht/senkrecht (w/s)" zweifach realisiert ist.

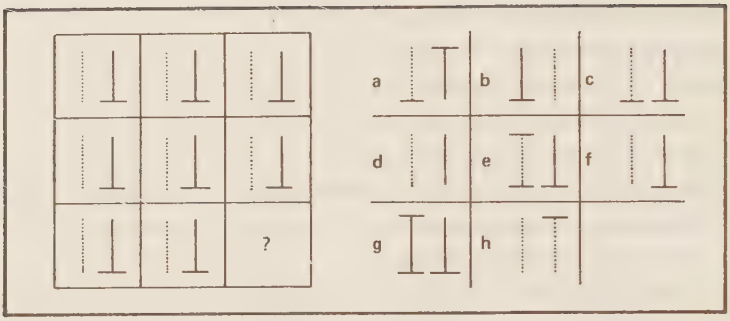


Abbildung 37 :
Gestaltstrategisch lösbares Item

Ähnlich wie bei den Aufgaben des SET A der Standard Progressive Matrices (RAVEN, 1958), ist hier das gesamte Schema als Muster auffaßbar, indem lediglich ein "Ausschnitt" (fehlerhaftes Musterelement) zu ersetzen ist. Diese "Gestaltstrategie" ist nicht Bestandteil der durch die Konstruktionsregeln/ Operationen definierten analytischen Problemlösefähigkeit.

Da nicht a priorisch auszuschließen war, daß sich in dem verbliebenen, homogenen Pool weitere "gestaltstrategisch" lösbare Aufgaben befinden, wurden alle 502 homogenen Items reinspiziert.

Dabei wurde deutlich, daß bei einer größeren Zahl von Items (76) die Strukturfestlegung geändert werden mußte, da es sich bei diesen Aufgaben um EINkomponenten -und nicht wie ursprünglich intendiert- um Zweikomponentenitems handelt. Zwar lassen sich, wie in u.a. Beispielen (Abb. 37), strukturanalytisch zwei Regeln isolieren (hier: Identität und Addition), doch ist durch gleichgerichtete Geltung (hier: linkes Item, RR:w/s: rechtes Item, RR:s) und vollständige Integration (RM:IR) eine Lösung allein dadurch möglich, daß EINE Regel (hier: Addition) auf die Gesamt-

figur angewendet wird. Dieses Problem tritt besonders häufig im Zusammenhang mit der "Identitätsregel" auf, weshalb es in Abbildung 38 abgekürzt als "Identitätsproblem" bezeichnet wird.

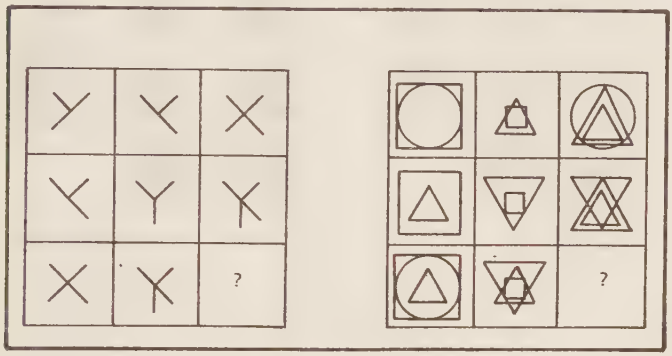


Abbildung 38 :
Beispiel für " Identitätsprobleme "

EINkomponentenitems sind jedoch, im Gegensatz zu den o.a. "gestaltstrategisch" lösbaren Items, vermutlich deshalb homogen, da sie, wie alle modellkonformen Aufgaben, nur analytisch lösbar sind.

Letztlich sind 5 modellunverträgliche, aber regelkonform konstruierte Aufgaben isolierbar, die mit hochkomplexen Repräsentanten realisiert wurden. Diese Konkretionen erfordern -im Vergleich zu allen anderen Matrizenitems- zusätzliche Wahrnehmungs- und Gedächtnisleistungen, die scheinbar nicht von allen Probanden erbracht werden und für die gemessene latente Denkfähigkeit nicht konstitutiv sind. Hier kann nur der allgemeine Hinweis erfolgen, Itemstrukturen "sparsam", mit möglichst wenigen Repräsentantenkombinationen herzustellen (s. Beispiel auf der nächsten Seite, Abbildung 39).

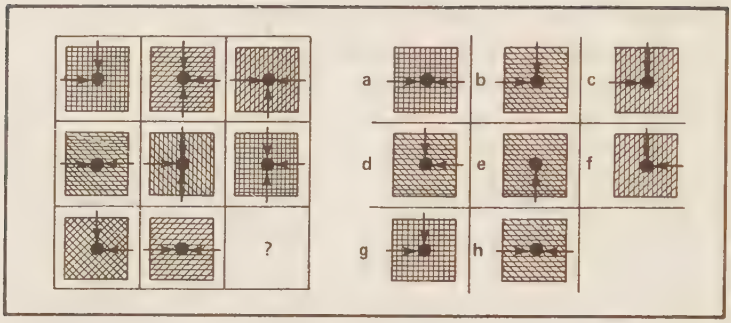


Abbildung 39 :
Beispiel für " Repräsentantenkomplexion "

Abbildung 40 (nächste Seite) gibt einen zusammenfassenden Überblick über die Ergebnisse, der in dem Kapitel dargestellten inhaltlich-strukturellen Analyse der nach Modellkontrollen zum RM selegierten Matrizenaufgaben und einer Reinspektion des gesamten Itemspools. Gestaltauflösungsprobleme, Konstruktionsfehler, IC-Fehler, gestaltstrategische Lösbarkeit und Komplexionen von Repräsentanten klären nahezu 3/4 aller beobachteten Iteminhomogenitäten. Damit ist das erklärte Ziel, möglichst allgemeine und auf das Konstruktionsrational beziehbare inhaltliche Deskriptoren für nicht-modellkonforme Aufgaben zu identifizieren, weitgehend erreicht. Zweifellos stellen sie zu diesem Zeitpunkt lediglich ad hoc - Erklärungen für mangelnde Modellkonformitäten dar, d.h. ihre empirische Gültigkeit ist in weiteren Konstruktionsexperimenten systematisch zu prüfen. Für etwa 1/4 (35) der 134 auffälligen Items konnten keine inhaltlichen Selektionsgründe identifiziert werden. Dies entspricht aber der Erwartung, da sich externe, durchführungsbedingte Störfaktoren, wie z.B. Speed-Einflüsse, sporadisch eben auch in Iteminhomogenitäten ausdrücken. Die Selektionsentscheidung kann daher für diese Aufgaben nur vorläufig sein, ihre meßtheoretische Brauchbarkeit soll in einer ergänzenden Untersuchung evaluiert werden (vgl. Kp. 12).

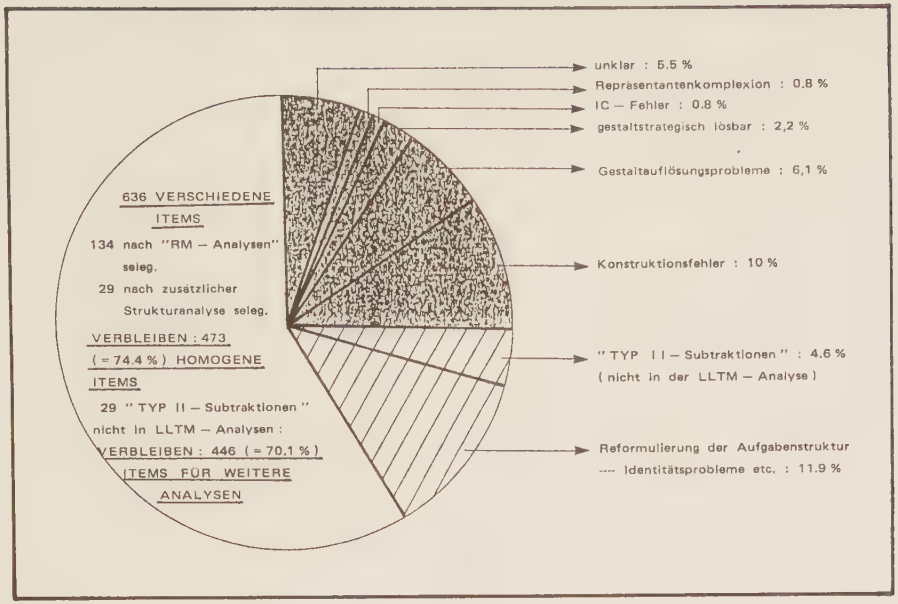


Abbildung 40 :
Ergebnisse der inhaltlich – strukturellen Reanalyse
des Aufgabenpools

In einer Kontingenzbetrachtung wurde der Frage nachgegangen, inwieweit Beziehungen zwischen verschiedenen Arten "technischer" und "qualitativer" Selektionskriterien bestehen. Ohne Schlüsse ziehen oder weiterreichende Interpretationen geben zu wollen, seien ihre Ergebnisse in Tabelle 6 wiedergegeben.

Modelltests	Inhaltliche Selektionsgründe						Signifikanzniveau
	1	2	3	4	5	6	
RES	36.4	66.7	35.7	80.0	40.0	11.4	p<.001
SCHULB	18.2	17.9	7.1	0.0	20.0	17.1	n.s.
MP	27.3	23.1	21.4	0.0	0.0	60.0	p<.01

Tabelle 6:

Relative Häufigkeiten, mit denen Items mit bestimmten qualitativen Selektionskriterien (1-6) in den verschiedenen Anpassungstests auffällig werden.

Hierbei:

1 = Konstruktionsfehler
2 = Gestaltauflösungsprobleme
3 = gestaltstrategische Lösbarkeit
4 = IC - Fehler
5 = Repräsentantenkomplexion
6 = unklar
RES = Residuenanalyse
SCHULB = Test auf Schulbildungsunabhängigkeit
MP = Tests durch multiple Partitionierungen

Aufgabenspezifische Inhomogenitätsquellen wurden in der vorliegenden Analyse gerade in sehr datennahen Residualanalysen nach WRIGHT und STONE (1979) sichtbar. Sie identifizieren fast 67% der

"Gestaltauflösungsprobleme" und 80% der "IC-Fehler". Die etwa in Analogie zur "klassischen" Konsistenzanalyse zu sehenden Anpassungskontrollen nach multiplen Partitionierungen der Probandenstichproben führen dagegen in erster Linie zu den nichtinhaltlich erklärbaren Aufgabenselektionen. Unter allen hier angebrachten Vorbehalten könnte man mutmaßen, daß dieser Modelltest eher sensitiv gegenüber externen, durch die Testadministration bedingten Verletzungen der Modellannahmen (z.B. Speed - Effekte) ist. Diese Vermutung wird auch durch die Beobachtung gestützt, daß aus dem offenbar "gespeedeten" Test 33 allein 7 Items nach diesem Verfahren selegiert werden mußten.

11.4 GÜTLIGKEITSKONTROLLEN ZUM LLTM

In Kapitel 4.2.1 wurde ein itemkonstruktionstechnologisch brauchbarer Operationsbegriff definiert. Es setzt voraus, daß HOMOGENE strukturelle Aufgaben gelöst werden, indem sachlogisch identifizierbare Regeln gesucht, erkannt und angewendet werden. Die zur richtigen Beantwortung eines Items notwendigen "Operationen über Regeln" sollen dabei seine Komplexität, ausgedrückt durch die b-Parameterschätzung des RM, erklären. Ob diese Annahmen über die Natur der Aufgabenbeantwortung bei den entwickelten und nach Modellkontrollen zum RM als hinreichend homogen ausgewiesenen Matrizenaufgaben haltbar, d.h. konsistent und verallgemeinerbar sind, soll nun geprüft werden, indem die LLTM-Schwierigkeitsparameter der angenommenen Operationen aus disjunkten Personen- und Itemstichproben geschätzt und die Modellanpassung mit einigen der in Kapitel 5.3 referierten Verfahren geprüft wird.

Um die Schwierigkeitsbeiträge der drei Realisierungsmodi, 8 Regeln und 3 Richtungen der Regelgeltung zu den Schwierigkeiten der verbliebenen 446 homogenen Items mit hinreichender Genauigkeit schätzen zu können, wurden die 35 Aufgabenstrukturmatrizen miteinander verketterter Testversionen so zusammengefaßt, daß 11 Strukturmatrizen mit Items aus je 3 Versionen und eine Matrix mit Aufgaben aus je 2 Versionen resultierten. Jede der 12 Strukturmatrizen umfaßt somit zwischen k=27 und k=48 (im Mittel k=39) Items. Dies entspricht in etwa den üblichen Itemmengen bisheriger Untersuchungen mit dem LLTM (vgl. SCHEIBLECHNER, 1972; FISCHER, 1974; SPADA, 1976; SPADA und KLUWE, 1981; WHITELY und SCHNEIDER, 1981; FISCHER und FORMANN, 1982; HORNKE und HABON, 1983, 1984a). Die Zusammenstellung von Strukturmatrizen dieses Umfangs folgt der Überlegung, daß jede Matrix einerseits ausreichend differenziert sein sollte, um stabile Schätzungen aller (14) Operationsparameter zu ermöglichen und andererseits nicht zu Konvergenzproblemen bei CML-Schätzungen der Item- und Operationsparameter führt (vgl. FISCHER und FORMANN, 1972).

In früheren Studien (s.o.) wurde davon ausgegangen, daß die Modellannahmen als gültig zu betrachten sind, wenn die Itemparameter des RASCH-Modells (b) durch die, nach Formel (10) aus den geschätzten Operationsparametern (η) rückgerechneten theoretischen Aufgabenschwierigkeiten (t) hinreichend genau approximierbar sind.

Datensatz	r(t,b)	m(t,b)	SD(b)	SD(t)
I	.77	.59	1.42	.99
II	.73	.52	1.25	.89
III	.80	.66	1.34	.98
IV	.67	.45	1.34	.93
V	.61	.38	1.29	.74
VI	.83	.65	1.44	1.13
VII	.78	.55	1.49	1.06
VIII	.74	.51	1.27	.88
IX	.84	.72	1.42	1.24
X	.82	.63	1.64	1.19
XI	.69	.39	1.34	.75
XII	.89	.80	1.56	1.46

Tabelle 7:

Produktmoment - Korrelationen "zusammengesetzter" Itemschwierigkeiten t und geschätzter Itemparameter b sowie deren Standardabweichung in 12 Datensätzen

Mittlere und hohe Korrelationen zwischen den RASCH-Itemparameterschätzungen (b) und den LLTM- Aufgabenschwierigkeiten (t) für jeden der 12 Datensätze lassen keineswegs den Schluß auf generell unterschiedliche Modellanpassungen in den "Teil-" Datensätzen zu, da sie empfindlich von der Variabilität der Itemschätzwerte abhängen. Dies sei ein Beispiel der Korrelation für den ersten Datensatz demonstriert:

r(t,b)	.77		
Testversion	1	2	3
r(t,b)	.84	.84	.63
SD(t)	1.09	1.09	.71
SD(b)	1.87	1.30	.93

Tabelle 8:
Korrelationen und Standardabweichungen innerhalb und über die drei Tests des Datensatzes I.

Die Korrelationen zwischen b- und t-Werten sind hoch, wenn diese stärker variieren, wie in der ersten und zweiten Testversion. Die Gesamtkorrelation wird aber durch die geringe Variabilität im dritten Test erheblich gemindert. Die direkte Gegenüberstellung der b- und t-Werte in graphischen Kontrollen (vgl. Anhang 3) zeigt denn auch, daß die Anpassung des linearen Modells jeweils lediglich für eine geringe Zahl von Aufgaben unzureichend ist. Gerade sehr "schwere" Aufgaben werden in ihrer Komplexität häufiger unterschätzt, während sehr "leichte" Items überschätzt werden. Gestützt wird diese Interpretation einerseits durch relativ niedrige Regressionskoeffizienten m(t,b), als auch durch die stets niedrigeren Standardabweichungen der t-Werte gegenuber denen der RASCH-Schwierigkeitsschätzungen b (vgl. Tabelle 7).

Üblicherweise wird versucht, eine verbesserte Modellanpassung durch Selektion solcher Aufgaben zu erzielen, für die das formulierte Lösungsmodell inadäquat zu sein scheint. Dabei wird jedoch die Erklärungskraft der zugrundegelegten Annahmen über die Lösungsoperationen für ALLE homogenen Aufgaben in Frage gestellt (s. GITTLER, 1983). Solange keine allgemeiner gültigen und in das Lösungsmodell integrierbaren Faktoren identifizierbar sind, die

systematisch die Schwierigkeiten jener abweichenden Aufgaben mitbedingen, scheint es sinnvoller, in diesem Sinne "problematische" Aufgaben beizubehalten und resultierende Beeinträchtigungen der Modellanpassung in den Teildatensätzen vorübergehend in Kauf zu nehmen.

Für die theoretische und praktische Bedeutsamkeit dieser Konstruktrepräsentation ist vielmehr entscheidend, ob es dennoch gelingt, interpretierbare und stabile, d.h. im vorliegenden Zusammenhang, über VERSCHIEDENE Personen- und Itemstichproben konstante Schätzungen der Operationsschwierigkeiten zu gewinnen. Das vorliegende Design ermöglicht erstmals, im Rahmen einer einzigen Untersuchung, mehrere Kreuzvalidierungen der Operationsschätzer in unabhängigen Item- und Personenstichproben (vgl. NÄHRER, 1980).

Betrachtet man jedoch vorab die Approximation der Schwierigkeitsschätzungen in Itemteilmengen, die GLEICHE Lösungsoperationen enthalten, so ist zunächst eine sehr gute Übereinstimmung der mittleren geschätzten und vorhergesagten Schwierigkeiten zu beobachten (vgl. Abb. 41, nächste Seite). Allerdings fällt auch hier wieder die geringere Varianz der t-Werte auf. Entgegen vorgängiger Erwartungen weisen die Schwierigkeitsmittelwerte in diesen Aufgabenuntermengen aber erhebliche Unterschiede auf, die vermutlich darauf zurückzuführen sind, daß durch vorgängige Itemselektionen die betreffende Operation in bestimmten Datensätzen vermehrt zusammen mit "leichteren" Operationen und in anderen Datensätzen vermehrt mit "schwereren" Operationen auftritt.

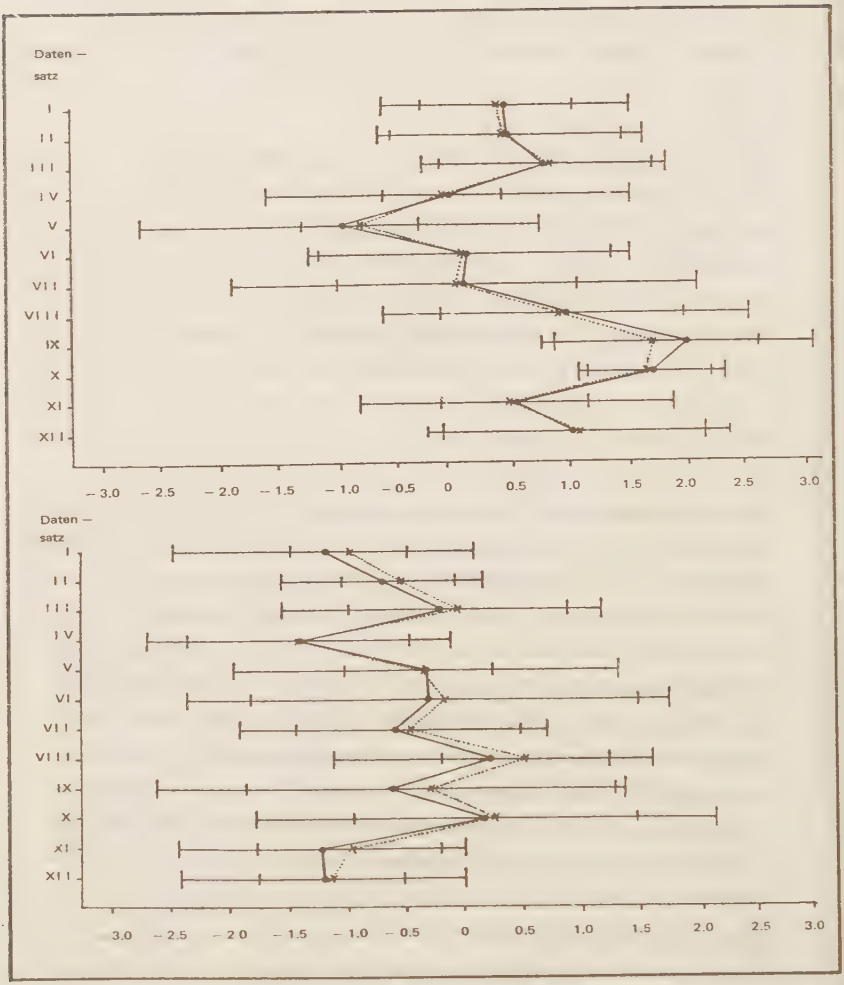


Abbildung 41 :
Arithmetische Mittel und Standardabweichungen geschätzter (• , I) und zusammengesetzter (X, |) Aufgabenschwierigkeiten bei Items mit der Operation " Variieren offener Gestalten " (oben) und bei Items mit der Operation " Identität " (unten)

In welcher Weise wirken sich nun die durch Über- bzw. Unterrepräsentation einzelner Operationen bedingten Unterschiede in den Schwierigkeitszusammensetzungen einzelner Itemteilmengen, auf die Schätzungen der 11 Operationsschwierigkeiten aus (3 Parameter (IC, Vg und w) wurden auf Null normiert, um die Schätzvoraussetzungen über den Rang der Strukturmatrix zu gewährleisten; vgl. FISCHER, 1974). Abbildung 42 (nächste Seite) gibt eine Überblick über die Verteilungen der 12 Schätzwerte für jede der "Realisierungsmodi", geltenden "Regeln" und "Richtungen der Regelgeltung" (vgl. auch Tabelle 9 in Anhang 4).

Vorliegende Hypothesen über relevante Komponenten der Problemlösung sind über Personen- und Aufgabenstichproben verallgemeinerbar, wenn sich die Schätzer eines Operationsparameters lediglich über sehr enge, nur durch die Schätzfehler (SF(n); Tabelle 9; Anhang 4) beschreibbare Konfidenzbereiche der Skala verteilen. Dies trifft für die Operationen "Seriation" und "Variation (g und o)" im wesentlichen zu. Bei den Operationen/ Regeln "Identität", "Subtraktion" und "Disjunktion" weisen einige Schätzer größere Abweichungen von einer dennoch deutlichen zentralen Tendenz auf, während für "Addition" und "Einzelelementenaddition" diese Tendenz nur schwach zu erkennen ist.

Verfolgt man die produktnormierten Operationsparameterschätzungen einzelner Datensätze in den Verteilungsdiagrammen, so ist festzustellen, daß die Abweichungen einzelner Schätzer eine Funktion des Umfangs der zugrundeliegenden Strukturmatrix bzw. ihrer Differenziertheit zu sein scheinen. Je undifferenzierter die Strukturzusammensetzung einer Matrix, d.h., je seltener einzelne Operationen im Lösungsprozeß der zusammengefaßten Itemmenge auftreten, desto eher weichen die zugehörigen Operationsschätzwerte von den jeweiligen Mittellagen der Verteilungen ab. Werden die Operationsschwierigkeiten hingegen aus umfangreicheren und differenzierteren Itemmengen geschätzt, so stellen sie in der Regel zufriedenstellende Näherungen der mittleren Operations-

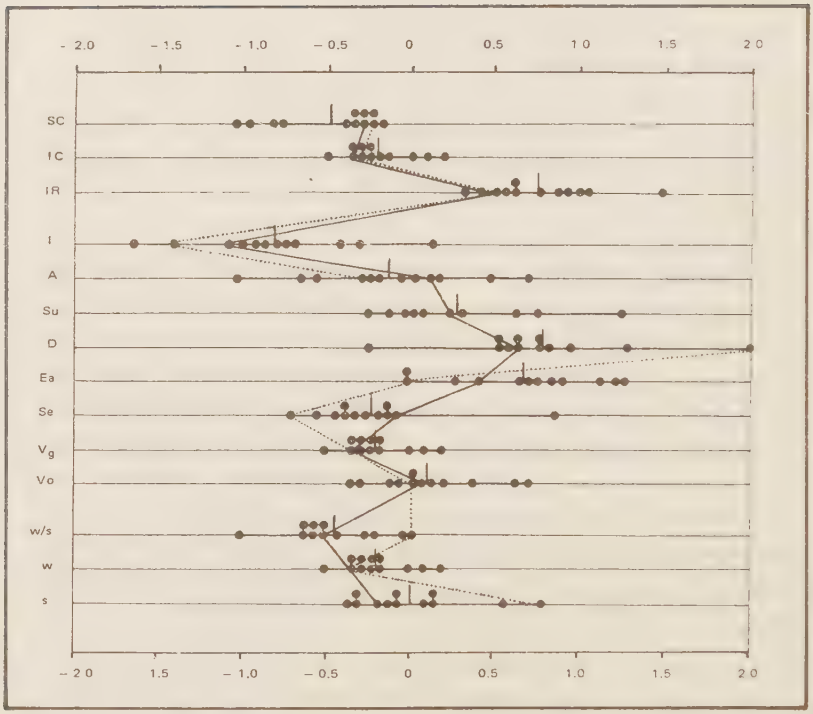


Abbildung 42 :
Verteilungen der 12 Schätzwerte für jede der RM , RE und RR.
Mittelwerte sind durch senkrechte Striche markiert.
Durch Verbindungslinien sind die Schätzwerte der kleinsten Itemmenge (k = 27 ; gepunktet) und der größten betrachteten Itemmenge (k = 48 Items ; durchgezogen) gekennzeichnet

schwierigkeiten dar. Als Beispiele für den erstgenannten Fall wurden in Abbildung 42 die Schätzwerte aus der "kleinsten" hier betrachteten Itemmenge (k=27) und für den letztenannten Fall die der "größten" Itemmenge (k=48) hervorgehoben.

SCHEIBLECHNER (1975) und SPADA (1976) weisen schon darauf hin, daß die Zuverlässigkeit und die Interpretierbarkeit der Operationsparameterschätzungen letztlich durch das Verhältnis der Aufgabenanzahl zur Anzahl der Operationen bestimmt ist. Um aufgabenspezifische Aspekte bei der Parameterschätzung zu vermeiden, empfehlen sich daher stets umfangreichere Aufgabenmengen, in denen die Operationen möglichst vollständig kombiniert sind. Obwohl aus der ursprünglich vollständig gekreuzten Strukturmatrix wegen mangelnder Aufgabenhomogenität und einigen nachträglich identifizierten Inkonsistenzen etwa 30% aller Strukturrealisationen eliminiert wurden, ist der gesamte verbliebene Pool von 446 homogenen und kontentvaliden Items als bei weitem ausreichend differenziert, um alle 14 angenommenen Basisoperationen zuverlässig schätzen zu können.

Um die Stabilisierung der Schätzwerte bei wachsendem Umfang der Strukturmatrix zu analysieren, wurden nunmehr die verbliebenen Items aus 6 bzw. 5 verketteten Testversionen zusammengefaßt. In die Analysen gehen nunmehr zwischen k=58 und k=85 (im Mittel k=74) strukturverschiedene Aufgaben ein.

Die theoretisch befriedigenden CML-Schätzungen konvergieren unter solchen Voraussetzungen, wenn überhaupt, nur sehr langsam. Die Operationsschwierigkeiten werden daher hier mittels multipler linearer Regressionen geschätzt, wobei die Schwierigkeitsschätzungen des RASCH-Modells als abhängige und die angenommenen Operationen als unabhängige Variablen eingingen. Um die praktische Vergleichbarkeit der Ergebnisse beider Schätzverfahren zu sichern, wurden zuvor die "Operationsgewichte" aus jenen 12 Teilmatrizen mittels multipler Regression (MR) geschätzt, die den

oben berichteten LLTM-Analysen zugrundelagen. Ein korrelations- und regressionsstatistischer Vergleich der "LLTM-" und "MR-"Operationsschätzwerte (vgl. HORNKE und HABON, 1985), zeigte eine für vorliegende Belange ausreichende Näherung der LLTM-Schätzer (η) durch die multiplen Regressionsgewichte (m). Wenn im folgenden MR-Operationsgewichte analysiert werden, so wird damit keineswegs eine prinzipielle Austauschbarkeit der Verfahren unterstellt. Vielmehr wird im vorliegenden Anwendungsfall dem rechentechnisch einfacheren Verfahren der MR der Vorzug gegeben, da es bei gleichem Ausgangsmaterial und ausreichend vielen annähernd normalverteilten Datenpunkten vergleichbare Ergebnisse erbringt. Die aus den nunmehr 6 zusammengefaßten Teilmatrizen geschätzten und anschließend produktnormierten Gewichte aller 14 Operationen (vgl. auch Tabelle 10, Anhang 5) sind in Abbildung 43 (nächste Seite) den jeweils darüber angeordneten MR-Schätzungen aus 12 Matrizen einander gegenübergestellt. Es ist unschwer zu erkennen, daß sich die Mittellagen der Schätzwertverteilungen zu jeder Operation (s. Tab. 10 und Markierungen in Abb. 43) nur unbedeutsam von den zentralen Tendenzen (arithmetischen Mitteln) bei den 12 Einzelschätzungen und den Schätzwerten unterscheiden, die schließlich aus der Gesamtmatrix (k=446 Items) gewonnen wurden. Die verbesserten Schätzvoraussetzungen führen hingegen zu deutlichen Verringerungen der Variablität der Schätzungen um die Mittelwerte der Verteilungen. Selbst bei den zuvor problematischen Operationen "Addition" und "Einzelelementenaddition" sind zentrale Schwierigkeitstendenzen durch die geringere Streuung der Schätzer nun deutlich bestimmt.

Die als Anpassungsmaße interpretierbaren multiplen Korrelationen R(b,m) zwischen den Itemparameterschätzungen (b) und den Operationsparameterschätzungen (m; s. Tabelle 10), nehmen in den etwa doppelt umfangreichen Datensätzen 1-6 durchschnittlich geringere Werte (AM(R)=.71) an, als in den Teildatensätzen I bis XII (AM(R)=.78 bzw. .76 für die LLTM-Schätzer). Eine geringfügige Verringerung des statistischen Zusammenhangs zwischen geschätzten

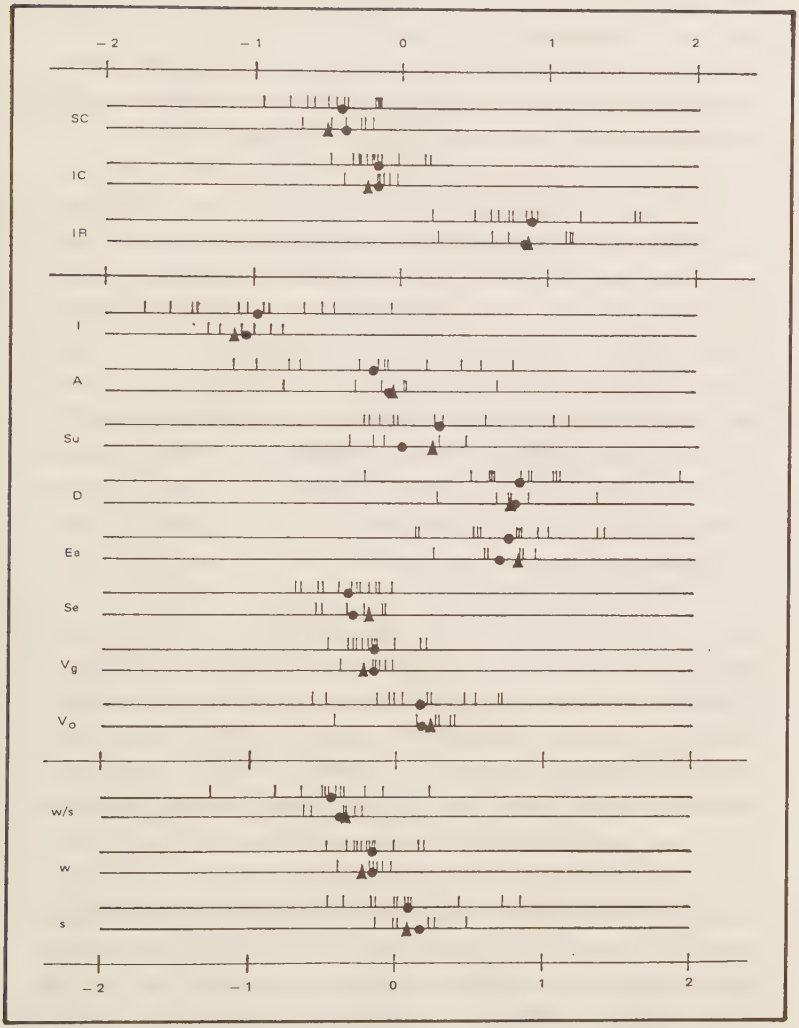


Abbildung 43 :
Vergleich von Operationsparameterschätzungen aus 12 (jeweils obere Linie) und 6 (jeweils untere Linie) zusammengefaßten Teilmatrizen . Punkte kennzeichnen die Mittelwerte der Verteilungen , Dreiecke die Schätzwerte aus k = 446 Items

und zusammengesetzten Aufgabenschwierigkeiten ist jedoch unter restriktiveren Bedingungen zu erwarten -die Zahl frei variierender Erklärungsparameter ist in den Sätzen 1-6 im Vergleich zu I-XII um die Hälfte reduziert- und spricht nicht prinzipiell gegen die Verallgemeinerbarkeit der unterstellten Annahmen über schwierigkeitserzeugende Komponenten der Aufgabenlösung.

Die empirischen Schwierigkeitsrelationen zwischen den unterstellten Operationen sprechen für die vorläufige Beibehaltung des Lösungsmodells, da sie weitgehend vorgängigen Erwartungen entsprechen. So weist die Operation "Identität", wie in allen Validierungen des ursprünglich von FORMANN (1973) entwickelten Vorläufermodells (vgl. PISWANGER, 1975; NÄHRER, 1977; HABON, 1981; HORNKE und HABON, 1983, 1984a) eine geringere Schwierigkeit auf, als alle anderen "geltenden Regeln". Ebenso bleibt die in den Arbeiten von FORMANN, PISWANGER und HORNKE und HABON vorgefundene Relation zwischen den Regeln "Addition" und "Variation (g)" erhalten.

Ähnliches zeigt sich bei den, in den Vorgängeruntersuchungen ebenfalls analysierten, Richtungen der Regelgeltung. Hier ist wieder die spaltenweise Orientierung gegenüber der zeilenweisen deutlich schwieriger, während nur ein geringer Schwierigkeitsunterschied zwischen den "RR:w" und "RR:w/s" besteht. Vorliegende zuverlässige Schätzungen der Schwierigkeitsbeiträge verschiedener Richtungen der Regelgeltung lassen das Finden und Durchführen einer Operation tendentiell leichter erscheinen, wenn die Regel sowohl in den Zeilen als auch in den Spalten des Aufgabenstamms gilt, als wenn sie NUR waagerecht auszuführen ist. Dies scheint inhaltlich plausibel, konnte aber bisher nur in den Analysen von NÄHRER (1977) und HORNKE und HABON (1983) gezeigt werden.

Auch die Komplexitäten erstmals untersuchter Regeln bzw. Operationen sind plausibel erklärbar. So zeigt sich beispielsweise, daß "Additionen" und "Disjunktionen" tendenziell "leichter" sind

als ihre jeweiligen Umkehroperationen "Subtraktion" bzw. "Einzelelementenaddition" (vgl. Tab. 10).

Allgemein scheint es schwerer zu sein, Regeln zu erkennen und anzuwenden, die sich auf einzelne FigurELEMENTE einer Gesamtfigur beziehen (D,EA), als solche, die über Veränderungen der Gesamtfigur einer Komponente definiert sind (I,Se,Vg oder Vo). "Additionen" bzw. "Subtraktionen" sind dagegen Mischformen, bei denen einmal Figurelemente und ein anderes mal die Figur selbst regelrelevante Merkmale darstellen.

Gerade im Hinblick auf die Konstruktion eines für CAT geeigneten Itempools, bestand ein wesentliches Ziel der Revision und Erweiterung des Konstruktionsrationals der Matrizenaufgaben darin, Operationen zu identifizieren, die es gestatten, "schwere", im oberen Skalenbereich lokalisierte Items zu entwickeln. Neben den Regeln "D" und "EA", sollte dies insbesondere durch zusätzliche "Suchoperationen" versucht werden, die dann erforderlich sind, wenn die beiden relevanten Figurkomponenten einer Aufgabe in abgestuft integrierter Form als "integrierte Komponenten" bzw. "integrierte Repräsentanten" dargeboten werden. In der Tat erschweren die beiden "Suchoperationen" die Lösung entsprechend realisierter Items und dies in vorhergesagter Weise. Die separierte Darstellung relevanter Figurenkomponenten erleichtert die Lösungsfindung, während ihre vollständige Integration eine schwierige Auflösungs- bzw. Suchoperation erfordert.

Durchschnittlich etwa zwei Drittel der Itemschätzwertvarianzen können durch die aus 12 Teildatenmengen geschätzten Schwierigkeiten von 14 unterstellten Lösungsoperationen aufgeklärt werden. Nurmehr 42% der GESAMTvarianz aller 446 Schwierigkeitsschätzer sind hingegen erklärt, wenn man die Operationsschätzungen aus der Gesamtmenge der Items zugrundelegt (vgl. Tab. 10, Anhang 5).

Die Gültigkeit des angenommenen linearen Lösungsmodells ist eingeschränkt, da es die Schwierigkeit von Aufgaben scheinbar extrem niedrigen und extrem hohen kognitiven Anforderungsniveaus tendentiell über- bzw. unterschätzt (vgl. Abb. 44, nächste Seite). Es ist nicht auszuschließen, daß Übungstransfer- oder Motivationseffekte während der Testbearbeitung die Schwierigkeiten gerade extremer Aufgaben beeinflussen. Da aber die Frage nach POSITIONSUNABHÄNGIGEN, d.h. von differentiellen Lern- oder Ermüdungseffekten unbeeinflußten Aufgabenschwierigkeitsschätzungen gerade für das variable adaptive Testen besonders bedeutsam ist, soll ihr in einer zusätzlichen Untersuchung nachgegangen werden.

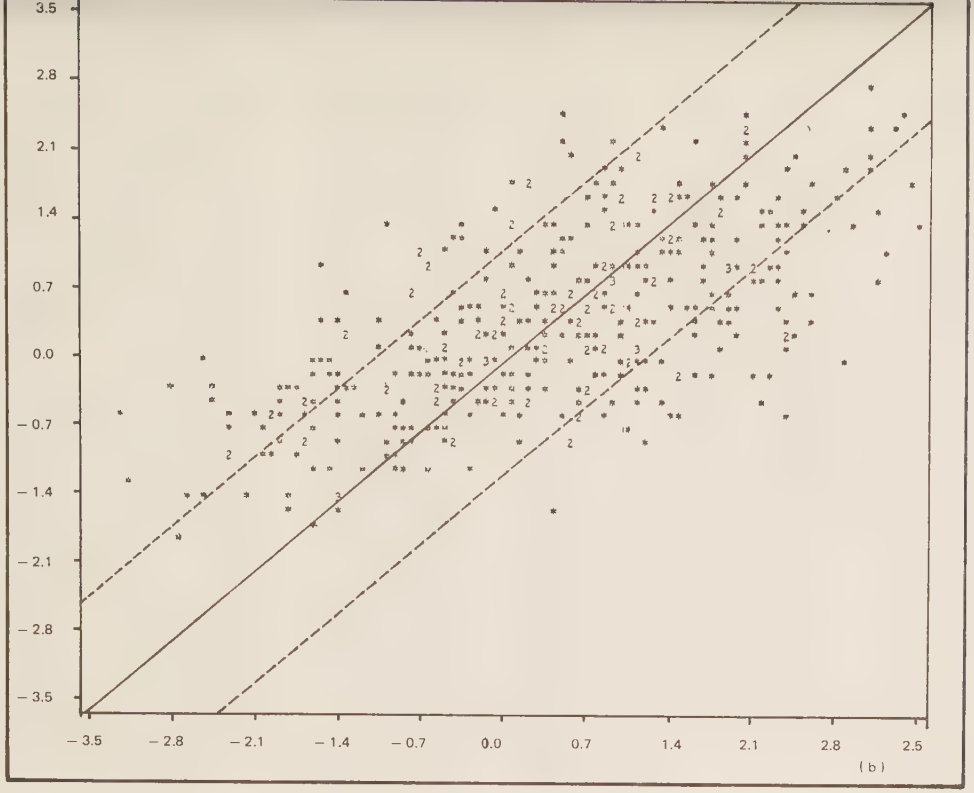


Abbildung 44 :
Vergleich der b — und t — Werte für alle k = 446 Items

12. EINE KREUZVALIDIERUNGSSTUDIE

SPADA (1976) und SCHEIBLECHNER (1975) weisen darauf hin, daß es nicht immer möglich ist, Tests mit übungsunabhängigen Aufgaben zu konstruieren (vgl. dazu auch FORMANN, KUBINGER und SCHUBERT, 1977). Es sind auch diagnostische Fragestellungen vorstellbar, bei denen gerade die Fähigkeit interessiert, einen Informationsgewinn aus Operationsdurchführungen flexibel auf strukturähnliche Aufgaben bzw. Problemstellung zu transferieren.

> "Für gewisse differentialpsychologische Fragen und Probleme der Eignungsmessung müßten gerade Tests mit Übungseinfluß von besonderer Bedeutung und praktischer Wichtigkeit sein, wo z.B. gerade die Lernfähigkeit gemessen werden sollte" (SCHEIBLECHNER, 1972, S.480).

Tests, in denen ein Übungstransfer/ Lernprozeß stattfindet, können durchaus Meßinstrumente bilden, die den oben gestellten Anforderungen genügen.

O.a. Autoren haben eine Reihe von Erweiterungen des LLTM vorgeschlagen, die es gestatten, bestimmte Formen des Übungstransfers während der Testbearbeitung unabhängig von der Zusammensetzung der Untersuchungsstichprobe, also spezifisch objektiv, zu quantifizieren.

Es wurde schon kurz darauf hingewiesen, daß beide Autoren zwischen "komponentenspezifischen Übungseffekten" (SCHEIBLECHNER, 1972) bzw. "operationsspezifischem Lernen" (SPADA, 1976) und "globalen Übungseffekten" / "globalem Lernen" unterscheiden. Im Gegensatz zum RM und LLTM, die Personen-, Aufgaben- bzw. Operationsparameter als Konstante annehmen, wird im linearen logistischen "Lernmodell" mit OPERATIONSSPEZIFISCHEM Übungstransfer die Schwierigkeit einer Operation in Abhängigkeit von der Übung

dieser Operation bei der selbständigen Bearbeitung vorangehender Aufgaben betrachtet.

Die lineare Nebenbedingung des Modells hat, in einer Version SPADA's (1976, S.145ff.), folgende Struktur:

$$(45) \qquad t_g = \sum_{j=1}^{m} f_{gj}(\eta_j - h_{gj}\,\beta_j) + c$$

wobei: f_{gj} ... Häufigkeit der Operation j in Item g
η_j ... Anfangschwierigkeit der Operation j
β_j ... Übungseffektparameter, der den maximal möglichen "Lerngewinn" bei der Operation j charakterisiert
h_{gj} ... (transformierte) Häufigkeit, mit der Operation j vor Aufgabe g geübt wurde
c ... Normierungskonstante

Zur Schätzung der Effektparameter β_j, für die bei Modellgeltung gilt: $0<\beta_j<\eta_j$, ist die Aufgabenstrukturmatrix F um eine hypothetische Übungsmatrix H=((f(gj)h(gj))) zu erweitern. Da h(gj) den ANTEIL des maximal erzielbaren Übungsgewinns β_j beschreiben soll, der bei Vorgabe des Items g durch Übung der Operation j in den Items u=1...g-1 erzielt wurde, muß die Summe der Operationsübungen bis zu diesem Zeitpunkt transformiert werden, sodaß $0 \leq \sum_{u=1}^{g-1} f_{uj} \leq 1$. Die von KEMPF (1974) vorgeschlagene logistische Transformation

$$(46) \qquad h_{gj} = \sum_{u=1}^{g-1} f_{uj} \;/\; 1 + \sum_{u=1}^{g-1} f_{uj}$$

formalisiert die Annahme, daß die erste Operationsübung die größte Abnahme der Operationsschwierigkeit zur Folge hat, während

der Effekt nachfolgender Übungen immer geringer wird. Zähler und Nenner des Ausdrucks (46) können noch mit einer Konstanten b multipliziert werden, die festlegt, wie schnell h(gj) gegen 1 bzw. $h_{gj}\,\beta_j$ gegen β_j als maximalen Übungsgewinn strebt. Abbildung 45 verdeutlicht den angenommenen Zusammenhang zwischen der Häufigkeit der Übung einer Operation und der Verringerung ihrer Schwierigkeit.

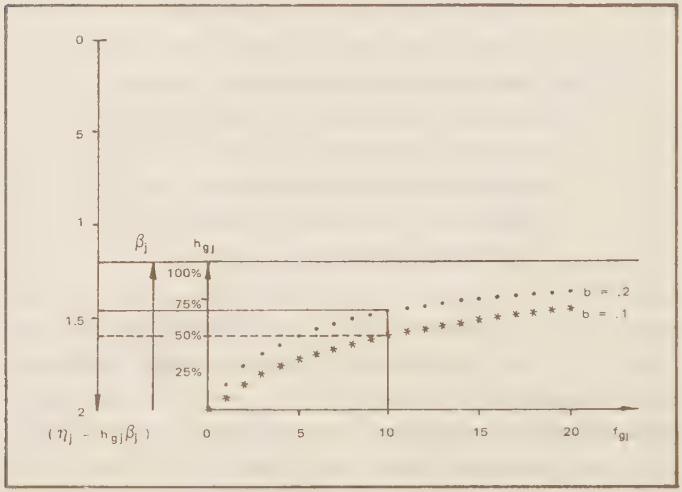


Abbildung 45 :
Zusammenhang zwischen maximalem Übungsgewinn β_j, dem bereits erreichten Übungsteil h_{gj}, der verbliebenen Restschwierigkeit $(\eta_j - h_{gj}\beta_j)$ und der Übungshäufigkeit f_{uj} bei $\eta_j = 2, \beta_j = .8$ und b = .1 bzw. b = .2; in modifizierter Form entnommen aus SPADA (1976)

Das Modell kann vereinfacht und die Zahl zu schätzender Parameter reduziert werden, wenn der Übungseffekt für alle Operationen gleich groß ist. Es ist dann $\beta_j = \beta$ für alle j.

Im Gegensatz zu diesem Modell, das einen Übungstransfer $h_{gj}\,\beta_j$ nur dann annimmt, wenn die Operation j in Aufgabe g ausgeführt werden muß, berücksichtigt das LLTM mit GLOBALEM Übungstrangsfer NICHT die Struktur f(gj) der Aufgabe g. Es wird vielmehr angenommen, daß die selbständige Bearbeitung einer Aufgabe und die Übung der darin auszuführenden Operationen zu einem generalisierten Effekt und damit zur Erhöhung der Lösungswahrscheinlichkeiten ALLER nachfolgenden Aufgaben führt. Dieser Übungsgewinn kann demnach sowohl durch die Senkung der Aufgabenschwierigkeit als auch durch die Erhöhung der latenten Fähigkeit der Personen ausgedrückt werden.

Dieser Exkurs über die formale Struktur linearer logistischer "Lernmodelle" verdeutlicht die sehr strengen Annahme darüber, wie der Übungstransfer während der Aufgabenbearbeitung stattfindet. Aus lerntheoretischer Sicht ist die Formalisierung von Lerngewinn auf Itemniveau in einem Modell problematisch, das auf der Annahme lokaler stochastischer Unabhängigkeit der Itembeantwortungen basiert. Der im LLTM abgebildete Übungstransfer auf eine Aufgabe k ist eine Funktion der über die Aufgaben 1...k-1 sumierten Operationshäufigkeiten. Es wird also unterstellt, daß ALLE Pbn vor Bearbeitung des Items k, k-1 Aufgaben tatsächlich bearbeitet und die zur RICHTIGEN Lösung führenden Operationen dabei geübt haben, UNABHÄNGIG davon, ob die Aufgaben schließlich richtig oder falsch beantwortet wurden. Der theoretisch interessantere reaktionskontingente Übungstransfer ist mit der Annahme lokaler stochastischer Unabhängigkeit unvereinbar (vgl. KEMPF, 1974).

Nur selten dürften theoretisch begründbare Hypothesen über Art und Umfang von Transfer- oder Motivationseffekten vorliegen, die mittels logistischer "Lernmodelle" formalisierbar wären. In der

Praxis wird man bestenfalls, d.h. bei einer hinlänglich großen Anzahl von Aufgaben, konkurrierende Varianten des dargestellten Modells anpassen können, um dann vielleicht das Modell mit der besten Anpassung beizubehalten. Abgesehen von der wissenschaftstheoretisch bedeutsamen Überlegung, ob nicht einem geringfügig ungünstiger angepaßten SPARSAMEN Modell ohne Übungseffektparameter der Vorzug gegeben werden sollte, läßt einerseits das theorielose Vorgehen, anderseits aber auch die Schwierigkeit einen Effektparameter für unterschiedliche Aufgaben und Reihenfolgen zu kreuzvalidieren, die vorgestellte Modellerweiterung problematisch erscheinen.

Liegen, wie dies ja in der Regel der Fall sein dürfte, keine theoretisch begründbaren Annahmen über den Einfluß von Übungseffekten über die Itemschwierigkeiten vor, sollte auf den Versuch solche Effekte zu quantifizieren verzichtet werden. Die Frage, ob Itemparameterschätzungen von der Position innerhalb eines Tests beeinflußt werden, kann hingegen durch ein spezifisches Testexperiment geprüft werden, in dem die Aufgabenreihenfolgen systematisch variiert und die Konstanz der resultierenden Schätzwerte kontrolliert werden.

12.1 DESIGN UND DURCHFÜHRUNG

Dem verbliebenen "homogenen" Matrizenitempool wurden 20 Aufgaben mit (maximal) unterschiedlichen Regel- bzw. Operationskombinationen entnommen. Da 20 Aufgaben selbstverständlich nicht ALLE möglichen Strukturkombinationen abdecken können, wurde die Auswahl auf Items mit "separierten Komponenten" beschränkt, die entweder "waagerecht und senkrecht" oder "waagerecht" gelten. Dabei wurde darauf geachtet, daß die 8 Regeln/ Operationen gleichhäufig (4-6 mal) in möglichst vielen verschiedenen Items anzuwenden sind. Der Aufgabenstrukturmatrix des experimentellen

Tests A (s. Tabelle 11) ist zu entnehmen, daß keine Aufgabe mit Regel- bzw. Operationswiederholung für die Nachuntersuchung ausgewählt wurde. Bei der Aufgabe mit der Positionsnr. 13 handelt es

	TEST A										TEST B
Item-Pos.	I	A	Su	D	Ea	Se	Vg	Vo	ws	w	Item-Pos.
*1	1	1	0	0	0	0	0	0	0	2	20
2	0	0	1	1	0	0	0	0	2	0	19
3	0	0	0	0	1	1	0	0	1	1	18
4	0	0	0	0	0	0	1	1	2	0	17
5	1	0	0	0	0	0	0	1	2	0	16
*6	0	1	0	0	0	0	1	0	2	0	15
*7	0	0	1	0	0	1	0	0	1	1	*14
8	0	0	0	1	1	0	0	0	0	2	13
9	1	0	0	1	0	0	0	0	2	0	9
10	0	1	0	0	0	1	0	0	2	0	10
11	0	0	1	0	0	0	0	1	2	0	11
12	0	0	0	0	1	0	0	1	2	0	12
13	0	1	0	0	0	0	0	0	0	1	8
14	0	0	1	1	0	0	0	0	0	2	7
*15	0	0	0	0	1	1	0	0	0	2	*6
16	0	0	0	0	0	0	1	1	0	2	5
17	1	0	0	0	0	0	0	1	1	1	4
18	0	1	0	0	0	0	1	0	0	2	*3
19	0	0	1	0	0	1	0	0	2	0	2
20	0	0	0	1	1	0	0	0	1	1	1

Tabelle 11:
Strukturmatrix und Positionen der 20 Aufgaben
in den Tests A und B

sich um ein Einkomponentenitem, dessen Strukturbeschreibung nachträglich modifiziert wurde (vgl. Kp. 11.3).

Obiger Tabelle ist ferner die veränderte Reihenfolge der 20 Aufgaben im zweiten experimentellen Test B zu entnehmen. Die letzten 8 Items des Tests A (Positionsnr. 13-20) bilden nun, in umgekehrter Reihenfolge, den ersten Testteil von B, während die ersten 8 Aufgaben von A (Positionnr. 1-8), ebenfalls in vertauschter Sequenz den Schlußteil von Test B bilden. Die Positionierung der Items im mittleren Testteil (9-12) bleibt unverändert. Wenn überzufällige Veränderungen der Aufgabenschwierigkeiten von der Position der Items im Test abhängen, so sollten die Schätzer für diese Items weitgehend stabil sein. Warming-up-Effekte, globaler oder operationsspezifischer Übungstransfer würden sich in vergleichbarer Weise auf die Schwierigkeiten dieser Aufgaben (9-12) auswirken, da die Items der vertauschten Testteile einen nahezu identischen strukturellen Aufbau aufweisen (vgl. Tab. 11).

Dieses Design läßt nicht nur die Beantwortung der hier vordringlich untersuchten Frage nach möglichen Positionseffekten zu, sondern erlaubt auch, zumindest für eine Stichprobe homogener Items zu prüfen, ob deren Lokalisationen auf der "Commonscale" (b(Pool); vgl. Kp. 11.2) valide, d.h. kontextunabhängig stabil sind. Wenn alle konstituierenden Annahmen des linearen logistischen Testmodells für die Lösung von Matrizenaufgaben hinlänglich gültig sind, so ist über die Stabilität der Itemparameterschätzungen hinaus zu erwarten, daß Schwierigkeitsunterschiede zwischen einzelnen Items auf lineare Kombinationen der Operationsschwierigkeiten (OP(Pool); vgl. Tabelle 10, Anhang 5) zurückführbar sind. Weiterhin sollten die aus den Daten der Testung A und der Testung B separat geschätzten Operationsschwierigkeiten OP(A) und OP(B) sowohl miteinander als auch mit den OP(Pool) hoch korreliert sein.

Einschränkend sei allerdings darauf hingewiesen, daß durch die relativ geringe Zahl von Aufgaben in den Tests A und B (k=20) die Schätzvoraussetzung für 9 Operationsparameter (ein Parameter (Vg) wurde auf Null normiert, s. Kp. 5.2) ungünstig sind, d.h. itemspezifische Effekte können die OP-Schätzer erheblich verzerren (vgl. Kp. 7).

Die Tests A und B wurden mit den Beispielaufgaben der Hauptuntersuchung im Frühjahr 1985 von zwei Stichproben Wehrpflichtiger (N(A)=398; N(B)=401) ohne Zeitbegrenzung bearbeitet. Die Zusammensetzung der Stichproben entspricht bezüglich Alter und Schulbildung den 32 Kalibrierungsstichproben der vorhergehenden Untersuchung.

12.2 ERGEBNISSE - DARSTELLUNG UND DISKUSSION

Item- und Personenparameter des RM wurden für Fitanalysen nach WRIGHT und STONE (1979) zunächst mittels PROX (vgl. Kp. 5) geschätzt. Die angestrebten Schätzwertvergleiche (s.o.) und die Beurteilung der Schätzwertkonstanz in disjunkten Teilstichproben hoch und niedrig scorender Probanden (Unterteilung der Stichproben an den Medianen der beiden Rohwertverteilungen; MDN (A;B)=10) und in Stichproben verschiedener Schulbildung ("Hauptschüler mit und ohne Abschluß" versus "Realschüler" und "Abiturienten") erfolgen hingegen auf der Grundlage von CML-Itemparameterschätzungen.

Gemessen an den Selektionskriterien der Hauptuntersuchung (vgl. 11.1.1) läßt keine der 799 Personen-Fitstatistiken t(v) Modellunverträglichkeiten erkennen, die eine Selektion entsprechender Antwortvektoren rechtfertigen würde. Damit steht die gesamte Datenmenge für die Schätzung und Beurteilung der Aufgaben- und Operationsparameter zur Verfügung.

Die Anpassungskontrollen zum RM weisen 2 Aufgaben (7 und 15 in A bzw. 14 und 6 in B) in BEIDEN Tests als nicht modellkonform aus. Beide Items wurden weder in den "technischen" Modellkontrollen der Hauptuntersuchung, noch in der anschließenden inhaltlichen Reinspektion der verbliebenen Itemmenge auffällig. In Kapitel 5.3 wurde schon darauf hingewiesen, daß alle verwendeten Anpassungsstatistiken selbst bei geringfügigen und praktisch vernachlässigbaren Modellunverträglichkeiten signifikante Werte annehmen, wenn die statistische Information über die Parameter hoch und damit die Konfidenzintervalle um ihre Schätzungen klein sind. In der Nachuntersuchung wurden die Aufgaben von etwa doppelt so vielen Pbn bearbeitet, wie in der Nachuntersuchung. Die mittlere Standardabweichung der Itemparameterschätzungen beträgt nurmehr SD(b(A/B))=.13, während sie bei den 502 Schätzwerten der Commonscale (vgl. Kp. 11.2) noch bei SD(b(Pool))=.20 liegt. Obgleich die damit verbundene Erhöhung der Teststärke gerade für die Differenzenstatistik u(g) von FISCHER und SCHEIBLECHNER (1970) zu einer unnötig strengen Beurteilung der Modellanpassung führt, soll hier konservativ verfahren und beide Items von weiteren Analysen ausgeschlossen werden. Darüber hinaus wurden drei weitere Aufgaben ausgeschieden, deren Fitstatistiken in jeweils einem der beiden Tests A und B knapp über den angesetzten kritischen Werten (t(crit)=5; u(crit)=2.56) liegen. Es handelt sich dabei um Aufgaben, die den, mitunter von Aufwärmeffekten betroffenen ersten Testdritteln (vgl. HABON, 1981) angehören (Item 1 und 6 in Test A und Item 3 in Test B).

Für die verbliebenen 15 Aufgaben wurden aus den Datensätzen A und B CML-Schätzungen der Item- und Operationsschwierigkeiten berechnet.

Abbildung 46 gibt einen Überblick über die angestellten Vergleiche und die ermittelten korrelativen Beziehungen.

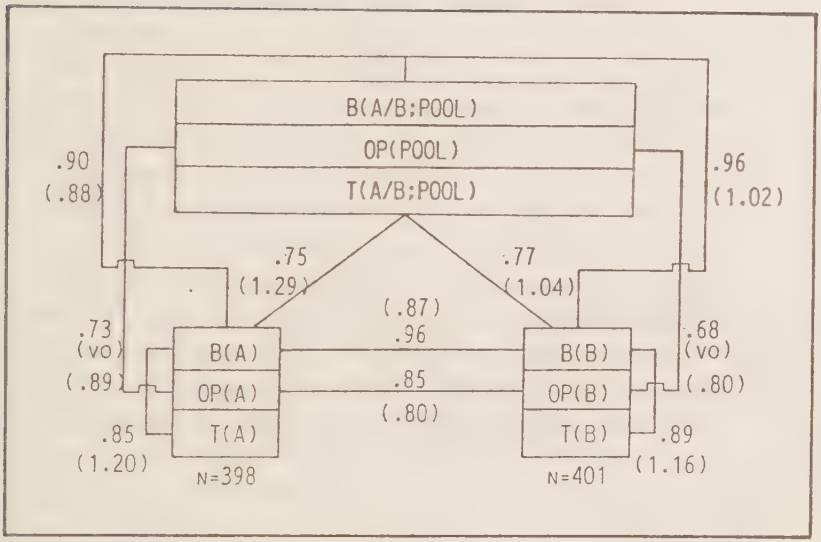


Abbildung 46 :
Überblick über die Ergebnisse der Kreuzvalidierungsstudie;
Produktmoment – Korrelationen und Regrssionskoeffizienten

Die aus VERSCHIEDENEN Pbn-Stichproben gewonnenen Schwierigkeitsschätzungen b(A) und b(B) der je 15 Aufgaben in Test A und B sind nahezu perfekt korreliert (Rangkorrelation r(b(A),b(B))=.96). Die Position dieser Items im Test ist somit OHNE nennenswerten Einfluß auf die Schätzungen ihrer Schwierigkeit im RM. Ähnlich hohe Korrelationen zwischen b(A) bzw. b(B) und den Lokalisationen der 15 Schätzer auf der Commonscale b(A/B;Pool) untermauern diese Aussage, da die Items in der Hauptuntersuchung in unterschiedlichen Testversionen auf verschiedensten Positionen angeordnet waren (Rangkorrelation r(b(A),b(A/B; Pool)=.87; r(b(B),b(A/B; Pool)=.98).

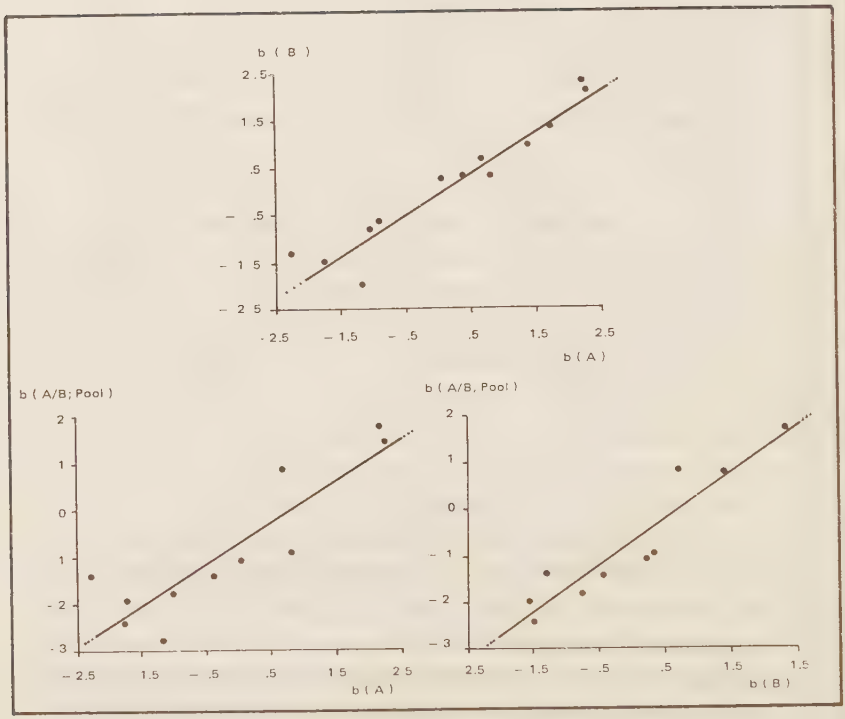


Abbildung 47 :
Graphische Gegenüberstellungen der b (A / B ; Pool) –, b (A) – und b (B) – Schätzungen

Unterstellt man einmal, daß diese Ergebnisse auf alle homogenen Aufgaben des Pools verallgemeinerbar sind, so ist zumindest EINE wesentliche Voraussetzung für seine Verwendung im Rahmen flexibler ADAPTIVER Teststrategien erfüllt. Vorgabe und Position einer Aufgabe sind hierbei antwortabhängig, d.h. eine bestimmte Aufgabe kann Personen mit verschiedenen Fähigkeitsausprägungen oder Personen gleicher Fähigkeit, aber unterschiedlich konsistentem Antwortverhalten (vgl. HORNKE, 1981, 1983), zu verschiedenen Zeitpunkten im Testprozeß administriert werden. Während alle Probanden, die einen konventionellen Test mit fester Aufgabenzahl und -reihenfolge bearbeiten, prinzipiell in gleicher Weise (positiv oder negativ) von evtl. Stellungseffekten betroffen sind, ist die positionsunabhängige Stabilität der Itemparameterschätzungen bei variablen adaptiven Tests -schon aus Gründen der Testfairness- eine notwendige Voraussetzung, da sonst die Schätzung der Personenparameter in unkontrollierter Weise mit den jeweiligen Testbedingungen konfundiert wäre.

Die strengste Kontrolle der formalen Konstruktrepräsentation, d.h. des LLTM und der Operationsparameterschätzungen, besteht darin, die Übereinstimmung zwischen den, aus einem neuen Datensatz geschätzten Schwierigkeiten b beliebiger homogener Items und den Schwierigkeitprognosen t zu prüfen. Letztere wurden hier auf der Grundlage der Strukturmatrix für die Tests A und B (vgl. Tab. 11) und den Operationsparameterschätzungen aus 446 Items berechnet (vgl. OP (Pool); Anhang 5, Tab. 10).

56 bzw. 59% der Variabilität der Aufgabenschwierigkeiten b(A) und b(B) werden durch die "theoretisch" vorhergesagten 15 Itemschwierigkeiten t(A/B; Pool) erklärt. Trotz der nach wie vor beträchtlichen nicht-erklärten Restvarianz, kann dieses Ergebnis zufriedenstellen, da der Anteil aufgeklärter Schwierigkeitsvarianz bei den Tests A und B um immerhin 14 bzw. 17% über dem, durch die OP(Pool) erklärten Anteil der Commonscale-Varianz liegt. Ähnlich hohe Korrelationen r(b(A),t(A/B;Pool)) und r(b(B); t(A/B;Pool))

waren zu erwarten, da b(A) und b(B) ihrerseits hoch interkorreliert sind.

Die aus den Daten des Test A und B geschätzten 9 Operationsparameter OP(A) und OP(B) sind dagegen kaum sinnvoll interpretierbar, da diesen Schätzungen lediglich 15 Aufgaben zugrundeliegen. So treten beispielsweise die Operationen 'Addition' und 'Variieren geschlossener Gestalten' lediglich zweimal bei der Lösung aller verbleibenden Items auf. Unter ähnlichen Bedingungen wurden schon früher erhebliche itemspezifische Abhängigkeiten der OP-Schätzungen beobachtet (vgl. Kp. 7). Die beobachteten hohen Korrelationskoeffizienten zwischen den Itemparameterschätzungen (b(A)/b(B)) und den rückgerechneten Aufgabenschwierigkeiten (t(A)/t(B)) belegen daher auch keinesfalls die Validität der Modellannahmen, sondern drücken, zumindest bis zu einem gewissen Grad, die Trivialität dieser Anpassung aus.

Die aus Abbildung 48 (nächste Seite) ersichtlichen Lokalisationen der aus Test A und B geschätzten Operationsschwierigkeiten verdeutlichen durch itemspezifische Effekte verzerrter Schätzwerte, besonders bei der Operation 'Vo' und den Richtungen der Regelgeltung 'w/s' und 'w'. Erstaunlicherweise bleibt trotz alledem die theoretisch postulierte Schwierigkeitsrangreihe für die Operationen 'I', 'Se', 'D', 'w/s' und 'w' erhalten (vgl. Kp. 11.4). Nimmt man die stark abweichenden Schätzwerte für die Operation 'Vo' aus der Betrachtung, so drückt sich diese "relative Stabilität" auch in mittelhohen korrelativen Beziehungen zu den Operationsschätzungen des gesamten Aufgabenpools (OP(Pool)) aus.

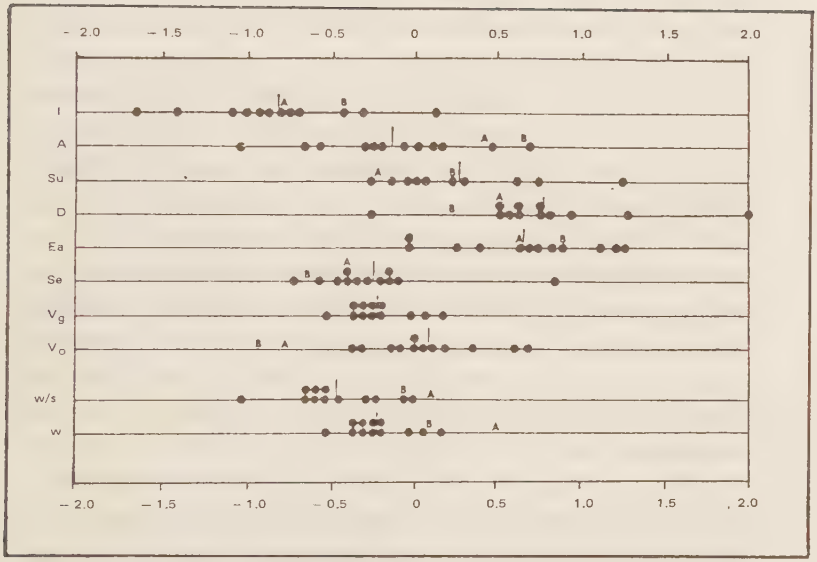


Abbildung 48 :
Lokalisationen der aus Test A und Test B geschätzten Operations – schwierigkeiten im Vergleich mit den Schätzwerten aus 12 Teildatensätzen der Hauptuntersuchung

13. SCHLUSSBEMERKUNGEN

Mit dem vorgeschlagegen Konzept der "rationalen Itemkonstruktion" konnten erstmals zwei der eingangs angesprochenen notwendigen Voraussetzungen für effiziente adaptiv - antwortabhängige Teststrategien geschaffen werden.

1. Der vorliegende Aufgabenpool hält für jede Variante konventionellen und flexiblen Testens ausreichend viele "RASCH-" homogene Problemlösungsaufgaben selbst extremer Leichtigkeit und Schwierigkeit bereit. So finden sich allein 68 strukturverschiedene Items mit Schwierigkeiten $b<-1.75$ und 57 Aufgaben mit $b>1.75$. In einem extrem breiten Skalenbereich stehen zwischen $b=-2.25$ und $b=2.25$ in neun gleichabständigen Intervallen jeweils mindestens 30 verschiedene Aufgaben zur Verfügung. Die Ergebnisse der Kreuzvalidierungsstudie deuten darauf hin, daß die Positionen der Items im Test ohne nennenswerten Einfluß auf ihre Schwierigkeitsschätzungen sind.

Weitgehend ungeklärt ist allerdings, in welcher Weise sich veränderte Präsentations- und Reaktionsbedingungen (Darstellung der Items auf einem Display, verschiedene Eingabemedien: Tastatur, Lichtgriffel, berührungsempfindlicher Bildschirm, Maus) auf die Homogenität und relative Schwierigkeit der Aufgaben auswirken. Eine entsprechende Studie, u.a. mit Matrizenaufgaben, die einigen dieser Aspekte nachgeht, wird derzeit unter der Leitung von Prof. Hornke (TH Aachen) durchgeführt. Im ungünstigen Fall wirken sich veränderte Vorgabebedingungen DIFFERENTIELL auf die Schätzungen verschiedener Items aus. Der Itempool wäre dann mit Daten aus computerunterstützten Testvorgaben zu rekalibrieren.

2. Das zugrundeliegende Konstrukt ist inhaltlich definiert und formal repräsentiert. Es existiert ein mathematisches Modell der Aufgabenbeantwortungen, das -für praktische Belange der Itemneu-

konstruktion und externen Validierung- mit hinlänglicher Stabilität die Homogenität und Schwierigkeit der Aufgaben als Funktion der Art und Komplexität hypothetischer Lösungsoperationen zu erklären vermag.

Zusätzliche Evidenzen über die Operationsannahmen erbrachte auch eine erste deskriptive Häufigkeitsanalyse der Antworten auf die ebenfalls nach rationalen Kriterien zusammengestellten Distraktoren. So entfallen beispielsweise auf Distraktoren, die das Ergebnis EINER richtigen Regelanwendung beinhalten, nahezu durchgängig mehr Antworten als auf Distraktoren, die mögliche Ergebnisse inkonsistenter oder nicht-intendierter Lösungsstrategien widerspiegeln. Dabei scheint die relative Häufigkeit einer Teillösung tendenziell von der in diese Teillösung eingehenden Operationskombinationen (geltende Regel und Richtung der Regelgeltung) abzuhängen.

Das zugrundeliegende "einfache" Bearbeitungsmodell berücksichtigt lediglich basale Lösungsoperationen, die sich auf das Suchen, Entdecken und Anwenden zuvor definierter (Konstruktions-) Regeln struktureller Aufgaben beziehen. Nur auf diese Weise ist es möglich ein SPARSAMES und damit empirisch validierbares Modell zu entwickeln, das zugleich ein KONSTRUKTIONSmodell für diese Art von Aufgaben ist. Die notwendige Beschränkung des Modells auf eine kleine Zahl beschreibender Parameter verbietet es, auch in den vorliegenden Untersuchungen aufgefallene item- oder itemgruppenspezifische Besonderheiten zu modellieren und empirisch zu prüfen. Solche "Spezifitäten" werden wohl insbesondere durch die NOCH NICHT objektivierte AUSWAHL einzelner graphischer Repräsentanten aus der "Repräsentantenliste" in die Itemkonstruktion hineingetragen. Eine durchaus denkbare und für die vollautomatisierte Konstruktion sogar notwendige Objektivierung dieses konstruktorischen Teilschrittes setzte die Entwicklung eines hochkomplexen Auswahl- und Kompositionssystems voraus, das wohl ebenso viele Regeln, wie Ausnahmen von diesen Regeln beinhalten

müßte. Es ist heute sehr wohl möglich, dieses kaum algorithmisch darstellbare "diffuse Wissen" der menschlichen Itemkonstrukteure bezüglich möglicher grahischer Repräsentationen der Konstruktionsregeln computergerecht zu repräsentieren (vgl. z.B. HAYES - ROTH, WATERMAN & LENAT, 1983) und so zu einer Art Expertensystem zu gelangen, das diese "Lücke" in einem ansonsten objektiven Konstruktionsprozeß schließen könnte. Für die konstruktbezogene Itemkonstruktion und Interpretation von Testwerten wäre allerdings wiederum nur dann etwas gewonnen, wenn es gelänge, im Sinne der formalen Konstruktrepräsentation, einen systematischen und differentiellen Einfluß dieser zusätzlichen (Realisierungs-) Regeln auf die Itemkomplexität empirisch aufzuzeigen.

Eines der dringlichsten und sicherlich schwierigsten Probleme aber besteht darin, die diagnostische Relevanz eines intern valide repräsentierten Konstrukts in einer Weise darzustellen, die auf wesentliche Konstituenten des Konstrukts (hier: basale Lösungsoperationen) Bezug nimmt.

Augenscheinlich wird ein Konstrukt etwa in prognostischen Situationen nützlich sein, wenn die durch das Konstrukt beschriebenen kognitiven Anforderungen in dem zu prognostizierenden Verhalten "auffindbar" sind. Die Auswahl eines Tests als Prognoseinstrument für ein interessierendes "Kriterium" würde nicht mehr allein mit irgendwelchen, einmal berichteten "Validitätskoeffizienten" des Tests begründet werden müssen, sondern könnte gezielt im Hinblick auf die kognitiven Anforderungen des erwarteten "Kriterienverhaltens" erfolgen. Eine solchermaßen theoriegeleitete Testdiagnostik ist auf der Basis einer einzelnen Konstruktrepräsentation nicht praktikabel. Vielmehr werden verschiedenste Aufgaben (i.w.S.) auf diese oder ähnliche Weise untersucht werden müssen. In experimentellen Generalisierbarkeitsstudien i.S. STERNBERG's "external validation" wird zu zeigen sein, ob und ggfls. welche der modellierten Komponenten über Itemformate und Gegenstandsbereiche verallgemeinerbar sind, um so möglicherweise einmal zu einem

System bekannter elementarer kognitiver Komponenten zu gelangen, das auch alltägliche und komplexe kognitive Leistungen beschreiben und erklären könnte.

14. Literaturverzeichnis

AMERICAN PSYCHOLOGICAL ASSOCIATION (APA) (1974):
Standards for educational and psychological tests,
American Psychological Association, Washington, D.C.

ANASTASI, A. (1967):
Psychology, psychologists and psychological testing,
Am. Psychologist, 22, 297-306

ANASTASI, A. (1985):
Some emerging trends in psychological measurement: A fifty-year perspective,
Applied Psychol. Meas., 9, 121-138

ANDERSEN, E.B. (1973):
Conditional inference and models for measuring,
Dissertation, Mental-hygiejnisk Forskningsinstitut, Kopenhagen

ANDERSEN, E.B. (1973b):
A goodness of fit test for the Rasch model,
Psychometrika, 38123-146

ANDERSEN, E.B. (1977):
Sufficient statistics and latent trait models,
Psychometrika, 42, 1, 69-82

ANDRICH, D. (1975):
The Rasch multiplicative binomial model: Applications to attitude data,
Res. Rep. 1, Measurement and Statistic Laboratory, Dept. of Education, Univ. of Western Australia

BAKER, F.B. (1984):
Technology and testing: State of the art and trends for the future,
J. of Educ. Meas., 21, 4, 399-496

BALLA, R.B. & MCDONALD, R.P. (1985):
Latent trait item analysis and facet theory - A useful combination,
Applied Psycol. Meas.,9, 191-198

BEJAR, I.I. (1983):
Introduction to item response models and their assumptions,
in: HAMBLETON, R.K. (Hrsg.):
Application of item response theory,
Educ. Res. Inst. of Brit. Columbia, Vancouver, B.C.

BEJAR, I.I. (1984):
Educatioal Diagnostic Assessment,
J. of Educ. Meas., 21, 175-189

BIRNBAUM, A. (1968):
Some latent trait models and their use in inferring an examinee's ability,
in: LORD, F.M. & NOVICK, M.R. (Hrsg.):
Statistical theories of mental test scores,
Addison-Wesley, Reading, Mass.

BLOOM, B.S. (Hrsg.) (1956):
Taxonomy of educational objectives. The classification of educational goals. Handbook I: Cognitive domain,
McKay, N.Y.

BLOOM, B.S. (Hrsg.) (1972):
Taxonomie von Lernzielen im kognitiven Bereich,
Beltz, Weinheim

BLOOM, B.S., HASTINGS, J.T. & MADAUS, G.F. (1971):
Handbook on formative and summative evaluation on student learning,
McGraw-Hill, N.Y.

BÜHLER, K. (1907):
Tatsachen und Probleme zu einer Psychologie der Denkvorgänge,
Archiv für die ges. Psychol., 12, 1-92

BUERLI, A. (1972):
Automatische MMPI - Systeme,
Schweizerische Zeitschrift für Psychol. und ihre Anwendungen, 31, 281-292

BUERLI, A. (1973):
Computerunterstützte Testdiagnostik,
Schweizerische Zeitschrift für Psychologie und ihre Anwendung, 1, 26-45

BUROS, O. (Hrsg.) (1965):
The sixth mental measurement yearbook,
Gryphon Press, Highland Park, N.J.

CAMPBELL, N.R. (1928):
Physics: The elements,
Cambridge University Press, London

CAMPBELL, D.T. & FISKE, D.W. (1959):
Convergent and discriminant validation by the multitrait - multimethod matrix,
Psychol. Bulletin, 56, 81-105

CAMPBELL, D.T. & STANLEY, J.C. (1963):
Experimental and quasi-experimental designs for research on teaching,
in: GAGE, N.L. (Hrsg.):
Handbook of research on teaching,
Rand McNally, Chicago

CARROLL, J.B. (1978):
How shall we study individual differences in cognitive abilities? - Methodological and theoretical perspectives,
Intelligence, 2, 87-115

CHOPPIN, B. (1968):
An item-bank using sample-free calibration,
Nature, 219, 870-872

CHOPPIN, B. (1976):
Recent developments in item banking,
in: DEGRUIJTER, D.N.M., VAN DER KAMP, TH. & CROMBAG, H.F. (Hrsg.):
Advances in psychological and educational measurement,
Wiley, N.Y.

CHOPPIN, B. (1984):
Computerized diagnostic testing,
J. of Educ. Meas., 21, 391-397

COOMBS, C.H., DAWES, R.M. & TVERSKY, A. (1975):
Mathematische Psychologie,
Beltz, Weinheim

CRONBACH, L.J. (1957):
The two disciplines of scientific psychology,
Am. Psychologist, 12, 671-684

CRONBACH , L.J. (1971):
Test validation,
in: THORNDIKE, R.L. (Hrsg.):
Educational measurement,
Am. Council of Education, Washington, D.C.

CRONBACH, L.J. & GLESER, G.C. (1965,2):
Psychological tests and personnel decisions,
Univ. of Illinois Press, Urbana, Ill.

CRONBACH, L.J. & MEEHL, P.E. (1955):
Construct validity in psychological tests,
Psychol. Bull., 52, 4, 281-302

DIVGI, D.R. (1981):
Does the Rasch model really work? Not if you look closely,
Paper presented at the annual meeting of the National Council for Measurement in Education, Los Angeles

DÖRNER, D. (1976):
Problemlösen als Informationsverarbeitung,
Kohlhammer, Stuttgart

DÖRNER, D. (1984):
Denken, Problemlösen und Intelligenz,
Psychologische Rundschau, 35, 1, 10-20

EBEL, R.L. (1962):
Content standard test scores,
Educ. Psychol. Meas., 22, 15-25

EMBRETSON (-WHITELY), S.E. (1977):
Models, meanings and misunderstandings:some issues in applying Rasch's theory,
J. of Educ. Meas., 14, 227-235

EMBRETSON (-WHITELY), S.E. (1980):
Multicomponent latent trait models for ability tests,
Psychometrika, 45, 479-494

EMBRETSON (-WHITELY), S.E. (1981):
Measuring aptitude processes with multicomponent latent trait models,
J. of Educ. Meas., 18, 67-84

EMBRETSON (-WHITELY),S.E. (1982):
Psychmetrics for test design,
Paper presented at the Invitational Conference on Item Response Theory and Computerized Adaptive Testing,
Minneapolis, Minn.

EMBRETSON (-WHITELY), S.E. (1983):
Construct validity: Construct representation versus nomothetic span,
Psychol. Bull., 93, 179-197

EMBRETSON (-WHITELY), S.E. (1985):
Test Design: Contributions from psychology, education and psychometrics,
Academic Press, N.Y.

EMBRETSON (-WHITELY), S.E & SCHNEIDER, L.M. (1981):
Information structure for geometric analogies: A test theory approach,
Applied Psychol. Meas., 5, 383-397

FISCHER, G.H. (Hrsg.) (1968):
Psychologische Testtheorie,
Huber, Bern

FISCHER, G.H. (1970):
Datenmodelle und Parametermodelle,
Zeitschrift für exp. und angew. Psychol., 17, 212-219

FISCHER, G.H. (1971):
Einige Gedanken über formalisierte psychologische Theorien,
Psychol. Beitr., 13, 3

FISCHER, G.H. (1972):
Conditional maximum-likelihood estimation of item parameters for a linear logistic test model,
Res. Bull. 9/72, Psychol. Inst. Univ. Wien

FISCHER, G.H. (1973):
The linear logistic test model as an instrument in educational research,
Acta Psychol., 37, 359-374

FISCHER, G.H. (1974):
Einführung in die Theorie psychologischer Tests,
Huber, Bern

FISCHER, G.H. (1976):
Some probabilistic models for measuring change,
in: DEGRUITER, D.N.M. & VAN DER KAMP , L.J.TH. (Hrsg.):
Advances in psychological and educational measurement,
Wiley, London

FISCHER, G.H. & FORMANN, A.K. (1972):
An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model,
Res. Bull. 11/72, Psychol. Inst. Univ. Wien

FISCHER, G.H. & FORMANN, A.K. (1982):
Some applications of logistic latent trait models with linear constraints on the parameters,
Applied Psychol. Meas., 6, 397-416

FISCHER, G.H. & PENDL, P. (1980):
Individualized testing on the basis ofthe dichotomous Rasch model,
in: VAN DER KAMP, L.J.TH., LANGERAK, W.F. & DEGRUIJTER, D.N.M.:
Psychometrics for educational debates,
Wiley, N.Y.

FISCHER, G.H. & SCHEIBLECHNER, H.H. (1970):
Algorithmen und Programme für das probabilistische Testmodell von Rasch,
Psychol. Beitr., 12, 24-51

FITZPATRICK, A.R. (1983):
The meaning of content validity,
Applied Psychol. Meas., 7, 3-13

FORMANN, A.K. (1973):
Die Konstruktion eines neuen Matrizentests und die Untersuchung des Lösungsverhaltens mit Hilfe des linearen logistischen Testmodells,
Phil. Diss. Univ. Wien

FORMANN, A.K. (1981):
Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch,
Zeitschrift für exp. und angew. Psychol., 4, 541-560

FORMANN, A.K., KUBINGER, K.D. & SCHUBERT, M.TH. (1977):
Matrizentests und ihre psychologischen Grundlagen - Eine testtheoretische Analyse,
Heilpädagogik, 20, 141-144

FORMANN, A.K. & PISWANGER, K. (1979):
Ein Rasch - skalierter sprachfreier Intelligenztest,
Beltz, Weinheim

FRICKE, R. (1972):
Über Meßmodelle in der Schulleistungsdiagnostik,
Schwann, Düsseldorf

FRICKE, R. (1974):
Kriteriumsorientierte Leistungsmessung,
Kohlhammer, Stuttgart

GARTNER, M.K. & STERNBERG, R.J. (1983):
Understanding general intelligence in terms of common reasoning components,
Unpublished Paper

GITTLER, G. (1983):
Konstruktion und Analyse neuer Raumvorstellungsaufgaben mittels probabilistischer Testmodelle,
in: LUER, G. (Hrsg.):
Bericht des 33. Kongresses der DGfPs in Mainz 1983

GLASER, R. (1963):
Instructional technology and the measurement of learning outcomes,
Am. Psychologist, 18, 519-521

GLASER, R. & NITKO, A.J. (1971):
Measurement in learning and instruction,
in: THORNDIKE, R.L. (Hrsg.):
Educational measurement,
American Council of Education, Washington, D.C.

GLASER, R. & PELLEGRINO, J.W. (1978):
Uniting cognitive process theory and differential psychology: Back home from the wars,
Intelligence, 2, 305-319

GOLDSTEIN, H. (1979):
Consequences of using the Rasch model for educational assessment,
Brit. Educ. Res. Journal, 5(2), 211-220

GREEN, B.F. (1970):
Comments on tailored testing,
in: HOLTZMAN, W.H. (Hrsg.):
Computer-assisted instruction, testing, and guidance,
Harper & Row, N.Y.

GUION, R.M. (1977):
Content validity - The source of my discontent,
Applied Psychol. Meas., 1, 1-10

GULLIKSEN, H. (1950):
Theory of mental tests,
Wiley, N.Y.

GUSTAFSSON, J.E. (1977):
The Rasch model for dichtomous items: theory, applications and a computer program,
Res. Rep. 63, Institute of Education, University of Göteborg

GUSTAFSSON, J.E. (1980):
Testing and obtaining fit of data to the Rasch model,
British J. of Mathematical and Statistical Psychol., 33, 205-233

GUSTAFSSON, J.E. (1981):
PML: a computer program for conditional estimation and testing the Rasch model for dichotomous items,
Res. Rep. 85, Institute of Education, University of Göteborg

GUTTMAN, L. & SCHLESINGER, J.M. (1967):
Systematic construction of distractors for ability and achievement test items,
Educ. and Psychol. Meas., 27, 569-580

HABON, M.W. (1981):
Sprachfreie Tests in der pädagogischen Diagnostik: Regelgeleitete Aufgabenkonstruktion durch formale Modelle,
Diplomarbeit, Universität Marburg

HÄUSSLER, P. (1981):
Denken und Lernen Jugendlicher beim Erkennen funktionaler Beziehungen,
Huber, Bern

HAMBLETON, R.K. et al. (1970):
A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple choice test,
J. of Educ. Meas., 7, 75-82

HAMBLETON, R.K. & Murray, L. (1983):
Some goodness of fit investigations for item response models,
in: HAMBLETON, R.K. (Hrsg.):
Applications of item response theory,
Educ. Res. Inst. of Brit. Columbia, Vancouver, B.C.

HAMBLETON, R.K. & NOVICK, M.R. (1973):
Toward an integration of theory and method for criterion - referenced tests,
J. of Educ. Meas., 10, 3, 159-170

HAMBLETON, R.K., ROBERTS, P.M. & TRAUB, R.E. (1970):
A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple choice test,
J. of Educ. Meas., 7, 75-82

HATTIE, J. (1985):
Methodology review: Assessing unidimensionality of tests and items,
Applied Psychol. Meas., 9, 139-164

HAYES-ROTH, F., WATERMAN, D.A. & LENAT, D.B. (1983):
Building expert systems,
Addison-Wesley, Reading, Mass.

HENDRICKSON, G.F. (1971):
The effect of differential option weighting on multiple choice objective tests,
J. of Educ. Meas., 8, 291-296

HERBIG, M. (1976):
Praxis lehrzielorientierter Tests,
Schwann, Düsseldorf

HILKE, R. (1980):
Grundlagen normorientierter und kriterienorientierter Tests,
Huber, Bern

HIVELY II, W., PATTERSON, H.L. & PAGE, S.A. (1968):
"Universe defined" system of arithmetic achievement tests,
J. of Educ. Meas., 5, 275-290

HOLTZMAN, W.H. (1971):
The changing world of mental measurement and its social significance,
Am. Psychologist, 26, 6

HORN, J.L. (1968):
On subjectivity in factor analysis,
J. of Educ. and Psychol. Meas., 27, 811-820

HORN, J.L. & KNAPP, J.R. (1973):
On the subjective character of the empirical base of Guilford's structure-of-intellect model,
Psychol. Bull, 80, 1, 33-43

HORNKE, L,F. (1975):
Vereinfachung und Demonstration eines sequentiellen Testverfahrens,
Zeitschrift für exp. und angew. Psychologie, 22, 47-54

HORNKE, L.F. (1976):
Grundlagen und Probleme antwortabhängiger Testverfahren,
Haag & Herchen, Frankfurt

HORNKE, L.F. (1976b):
Simulation einer Zufalls-Item x Probanden Matrix,
Zeitschrift für exp. und angw. Psychologie, 23, 230-239

HORNKE, L.F. (1977):
Antwortabhängige Testverfahren: Ein neuartiger Ansatz psychologischen Testens,
Diagnostica, 23, 1-14

HORNKE, L.F. (1978):
Vergleich zweier adaptiv-antwortabhängiger Teststrategien,
Diagnostica, 24 103-112

HORNKE, L.F. (1981):
Computerunterstütztes Testen (CUT) von Prüfungsangst,
Zeitschrift für Diff. und Diagn. Psychol., 2, 325-335

HORNKE, L.F. (1983):
Computerunterstütztes Testen - Eine bewertende empirische Untersuchung,
Zeitschrift für Diff. und Diagn. Psychol., 4, 232-344

HORNKE, L.F. & HABON, M.W. (1983):
Evaluation einer rationalen Itemkonstruktion,
in: OLECHOWSKI, R. (Hrsg.):
Der Beitrag der empirischen Erziehungswissenschaft zur Praxisverbesserung von Schule, Unterricht und Erziehung,
Braunschweiger Studien zur Erziehungs- und Sozialwissenschaft, Bd. 10, Braunschweig

HORNKE, L.F. & HABON, M.W. (1984a):
Erfahrungen zur rationalen Konstruktion von Testaufgaben,
Zeitschrift für Diff. und Diagn. Psychol., 3, 203-212

HORNKE, L.F. & HABON, M.W. (1984b):
Regelgeleitete Konstruktion und Evaluation von nicht - verbalen Denkaufgaben,
Wehrpsychologische Untersuchungen, Heft 4

HORNKE, L.F. & HABON, M.W. (1985):
Kognitive Lösungsoperationen bei regelgeleitet konstruierten figuralen Denkaufgaben,
Arbeitsbericht AL-2-85 des PsychDstBw, Bonn

HOTELLING, H. (1933):
Analysis of a complex of statistical variables into principal components,
J. of Educ. Psychol., 24, 417-441, 498-520

JACOBS, P.J. & VANDEVENTER, M. (1972):
Evaluating the teaching of intelligence,
J. of Educ. and Psychol. Meas., 32, 235-284

JENSEMA, C.J. (1977):
Bayesian tailored testing and the influence of item bank characteristics,
Applied Psychol. Meas., 1, 111-120

JÖRESKOG, K.G. (1969):
A general approach to confirmatory maximum likelihood factor analysis,
Psychometrika, 34, 183-202

KEMPF, W.F. (1974):
Drei Arbeiten über verallgemeinerte Rasch-Modelle,
IPN-Arbeitsbericht Nr.10, Kiel

KIRCHENKAMP, TH.P. (1984):
Kognitionspsychologische Modelle und rationale Itemkonstruktion,
Unveröffentl. Diplomarbeit, Univ. Düsseldorf

KLAUER, K.J. (1972):
Einführung in die Theorie lehrzieorientierter Tests,
in: K.J. KLAUER et al. (Hrsg.):
Lehrzielorientierte Tests,
Schwann, Düsseldorf (3. Aufl. 1975)

KLAUER, K.J. (1978):
Kontentvalidität,
in: K.J. KLAUER (Hrsg.):
Handbuch der pädagogischen Diagnostik,
Schwann, Düsseldorf

KLAUER, K.J. (1984):
Kontentvalidität,
Diagnostica, 30, 1-23

KLUWE, R.H. & SPADA, H. (1981):
Zwei Modelle der Denkentwicklung und ihr Bezug zur Theorie von Piaget,
in: KLUWE, R.H. & SPADA, H. (Hrsg.):
Studien zur Denkentwicklung,
Huber, Bern

KRATHWOHL, D.R., BLOOM, B.S. & MASIA, B.B. (1964):
Taxonomy of educational objectives. The classification of educational goals,
Handbook II: Affective Domain,
McKay, N.Y.

KRAUTH, J. & LIENERT, G.A. (1973):
Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin: Ein multivariates nichtparametrisches Verfahren zur Aufdeckung von Typen und Syndromen,
Alber, Freiburg

LAZARSFELD, P.F. (1950):
The logical and mathematical foundation of latent structure analysis,
in: STOUFFER, S.A. et al. (Hrsg.):
Measurement and prediction,
Princeton Univ. Press, Princeton, N.J.

LIENERT, G.A. (1969/3):
Testaufbau und Testanalyse,
Beltz, Weinheim

LOEVINGER, J. (1957):
Objective tests as instruments of psychological theory,
Psychol. Rep., 3, 635-694

LORD, F.M. (1970):
Some test theory for tailored testing,
in: HOLTZMAN, W.H. (Hrsg.):
Computer assisted instruction, testing and guidance,
Harper & Row, N.Y.

LORD, F.M. (1977):
A broad-range tailored test of verbal ability,
Applied Psychol. Meas., 1, 95-100

LORD, F.M. (1980):
Applications of item response theory to practical testing problems,
Erlbaum, Hillsdale

LORD, F.M. (1983):
Estimating the imputed social cost of errors of measurement,
Educational Testing Service, Princeton, N.J.

LORD, F.M. & NOVICK, M.P. (1968):
Statistical theories of mental test scores,
Addison-Wesley, Reading, Mass.

LUMSDEN, J. (1976):
Test theory,
Annual Rev. of Psychol., 27, 251-280

LUMSDEN, J. (1978):
Tests are perfectly reliable,
Brit. J. of Stat. Psychol., 38, 19-26

MARTIN-LöF, P. (1973):
Statistika Modeller,
Institut for Forsakringsmatematik och Matematik Statistik,
Univ. Stockholm

MCNEMAR, Q. (1964):
Lost: Our intelligence? Why?
Address of the President to the 72. Annual Convention of the APA 1964, Los Angeles

MEAD, R.J. (1975):
Analysis of fit to the Rasch model,
Doctoral dissertation, Univ. of Chicago

MERZ, F. & KALVERAM, K.TH. (1965):
Kritik der Differenzierungshypothese der Intelligenz,
Arch. ges. Psychol., 117, 287

MESSICK, S. (1972):
Beyond structure: In search of functional models of psychological process,
Psychometrika, 37, 357-375

MESSICK, S. (1975):
The standard problem: Meaning and values in measurement and evaluation,
Am. Psychologist, 30, 955-966

MESSICK, S. (1981):
Constructs and their vicissitudes in educational and psychological measurement,
Psychol. Bull., 89, 575-588

MILL, J.ST. (1885):
Systeme der deduktiven und induktiven Logik,
Gesammelte Werke, Bd. 2,
Fues's Verlag, Leipzig

MILLMAN, J. & ARTER,J.A. (1984):
Issues in item banking,
J. of Educ. Meas., 21, 4, 315-330

MISPELKAMP, H.B. (1985):
Theoriegeleitete Sprachtestkonstruktion,
Lang, Bern

MOOSBRUGGER, H. (1983):
Neuere psychologische Testtheorien und ihre Anwendung in der Psychodiagnostik,
Arbeiten aus dem Institut für Psychologie der Univ. Frankfurt/Main

MULHOLLAND, T., PELLEGRINO, J.W. & GLASER, R. (1980):
Components of geometric analogy solution,
Cognitive Psychology, 12, 252-284

NÄHRER, W. (1977):
Modellkontrollen bei der Anwendung des linearen logistischen Testmodells in der Psychologie,
Phil. Diss. Univ. Wien

NÄHRER, W. (1980a):
Modellkontrollen bei der Anwendung des linearen logistischen Testmodells,
Diagnostica, 26, 2, 112-118

NÄHRER, W. (1980b):
Zur Analyse von Matrizenaufgaben mit dem linearen logistischen Testmodell,
Zeitschrift für exp. und angew. Psychol., 27, 4, 553-564

NEWBOULD, C.A. & MASSEY, A.J. (1977):
A computerized item banking system,
Brit. J. Educ. Technology, 8, 113-123

NEWELL, A. & SIMON, H.A. (1972):
Human problem solving,
Prentice Hall, N.J.

NEYMAN, J. & SCOTT, E.L. (1948):
Consistent estimates based on partially consistent observations,
Econometrica, 16, 1-32

NUSSBAUM, A. (1980):
Konstruktion, Planung und Analyse lehrzielorientierter Tests auf der Grundlage der Generalisierbarkeitstheorie,
Phil. Diss. TH Aachen

ORTH, B. (1974):
Einführung in die Theorie des Messens,
Kohlhammer, Stuttgart

OSBURN, H.G. (1968):
Item sampling for achievement testing,
J. of Educ. and Psychol. Meas., 28, 95-104

PATNAIK, D. & TRAUB, R.E. (1973):
Differential weighting by judged degree of correctness,
J. of Educ. Meas., 10, 281-286

PEARSON, K. (1901):
On lines and planes of closest fit to a system of points in space,
Philosophical Magazine, 2, 557-572

PELLEGRINO, J.W. & LYON, D.R. (1979):
The components of a componential analysis,
Intelligence, 3, 169-186

PISWANGER, K. (1975):
Interkulturelle Vergleiche mit dem Matrizentest von Formann,
Phil. Diss. Univ. Wien

PUTZ-OSTERLOH, W. (1981):
Problemlösungsprozesse und Intelligenztestleistung,
Huber, Bern

RASCH, G. (1960):
Probabilistic models for some intelligence and attainment tests,
Nielson & Lydiche, Kopenhagen

RASCH, G. (1966):
An informal report on a theory of objectivity in comparisons,
Proceedings on the NUFFIC international summer session in science at "Het Oude Hof", Den Haag

RAVEN, J.C. (1956):
Standard Progressive Matrices,
Lewis & Co, London

RAVEN, J.C. (1962):
Advanced Progressive Matrices, Set II,
Lewis & Co, London

REILLY, R.R. & JACKSON, R. (1973):
Effects of empirical option weighting on reliability and validity of an academic aptitute test,
J. of Educ. Meas., 10, 185-194

REINERT, G., BALTES, P.B. & SCHMIDT, L. (1965):
Faktorenanalytische Untersuchungen zur Differenzierungshypothese der Intelligenz: Die Leistungsdifferenzierungshypothese,
Psychol. Forschung, 28, 246-300

RENTZ, R.R. & BASHAW, W.L. (1975):
Equating reading tests with the Rasch model,
Educational Resource Laboratory, Athens, Georgia

RESNICK, L.B. (1976):
The nature of intelligence,
Erlbaum, Hillsdale

REVENSTORF, D. (1976):
Lehrbuch der Faktorenanalyse,
Kohlhammer, Stuttgart

ROID, G.H. & HALADYNA, T.M. (1982):
A technology for testitem writing,
Academic Press, N.Y.

RUDINGER, G. & RUPPEL, H. (1981):
Quantitative Veränderungen in der kognitiven Entwicklung: Beschreibung, Erfassung und theoretische Erklärung,
in: KLUWE, R.H. & SPADA, H. (Hrsg.):
Studien zur Denkentwicklung,
Huber, Bern

SAMEJIMA, F. (1969):
Estimation of latent ability using a response pattern of graded scores,
Psychometric Monograph Supplement, 17

SCHEIBLECHNER, H.H. (1972):
Das Lernen und Lösen komplexer Denkaufgaben,
Zeitschrift für exp. und angew. Psychol., 19, 476-505

SCHEIBLECHNER, H.H. (1974):
Die Sozialstruktur großer Gruppen,
in: KEMPF, W.F. (Hrsg.):
Probabilistische Modelle in der Sozialpsychologie,
Huber, Bern

SCHEIBLECHNER, H.H. (1975):
Kritik der Anwendung des linearen logistischen Modells in der Psychologie,
Ber. 29. Kongreß der DGfPs 1974,
Hogrefe, Göttingen

SCHEIBLECHNER, H.H. (1980):
Tests und Testtheorie,
in: ASANGER, R. & WENNINGER, G. (Hrsg.):
Handwörterbuch der Psychologie,
Beltz, Weinheim

SCHOTT, F. (1973):
Anwendungsmöglichkeiten einer Matrix aus zweidimensionalen Aufgabenklassen in der psychologischen Therapie,
Psychol. Praxis, 17, 125-136

SCHOTT, F. (1983):
Probleme kriteriumsorientierter Leistungsmessung: Zum praktischen Nutzen lehrzielorientierter Tests im Unterricht,
in: HORN, R. et al. (Hrsg.):
Tests und Trends. 3. Jahrbuch der Pädagogischen Diagnostik,
Beltz, Weinheim

SCHOTT, F. & WIEBERG, H.-J. (1984):
Regelgeleitete Itemkonstruktion,
Diagnostica, 30, 47-66

SHOEMAKER, D.M. (1973):
Priciples and procedures of multiple matrix sampling,
Ballinger, Mass.

SIGEL, I.E. (1963):
How intelligence Tests limit understanding of intelligence,
Merrill-Palmer Quarterly 9, 30-56

SIMON, H.A. (1979):
Information processing models of cognition,
Annual Review of Psychol., 30, 363-396

SPADA, H. (1969a):
Grundlagenforschung zur Intellingenzdiagnostik,
Phil. Diss. Univ. Wien

SPADA, H. (1969b):
Invariante Intelligenzdimensionen als Basis für Selektion und Klassifikation,
in: IRLE, M. (Hrsg.):
Ber. 26. Kongreß der DGfPS in Tübingen 1968,
Hogrefe, Göttingen

SPADA, H. (1973):
Denk- und Lernmodelle der Rasch-Meßtheorie unter psychologischen, formalen und didaktischen Aspekten,
in: SPADA, H., HÄUSSLER, P. & HEYNER, W. (Hrsg.):
Denkoperationen und Lernprozesse als Grundlage für lernerorientierten Unterricht,
IPN - Arbeitsberichte, 5, Kiel

SPADA, H. (1976):
Modelle des Denkens und Lernens,
Huber, Bern

SPADA, H. & FREY, K. (1973):
Die Formulierung von Lernzielen und die Konstruktion lernzielorientierter Tests mit Hilfe formaler Modelle,
in: SPADA, H., HÄUSSLER, P. & HEYNER, W. (Hrsg.):
Denkoperationen und Lernprozesse als Grundlage für lernerorientierten Unterricht,
IPN - Arbeitsberichte, 5, Kiel

SPADA, H., HÄUSSLER, P. & HEYNER, W. (Hrsg.):
Denkoperationen und Lernprozesse als Grundlage für lernerorientierten Unterricht,
IPN - Arbeitsberichte, 5, Kiel

SPEARMAN, C. (1904):
'General intelligence', objectively determined and measured,
Am. J. Psychol., 15, 201-293

STELZL, I. (1979):
Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten,
Zeitschrift für exp. und angew. Psychol., 14, 652-672

STERNBERG, R.J. (1977):
Intelligence, information processing, and analogical reasoning:
The componential analysis of human abilities,
Erlbaum, Hillsdale

STERNBERG, R.J. (1979):
The nature of mental abilities,
Am. Psychologist, 33, 214-230

STERNBERG, R.J. (1980):
Scetch of a componential subtheory of human intelligence,
The Behav. and Brain Sciences, 3, 573-614

STERNBERG, R.J. (1981):
Nothing fails like success: The search for an intelligent paradigm for studying intelligence,
J. of Educ. Psychol., 73, 2, 142-155

STERNBERG, R.J. & TURNER, M.E. (1981):
Components of syllogistic reasoning,
Acta Psychologica, 47, 245-265

SUPPES, P. & MORNINGSTAR, M. (1972):
Computer-assisted instruction at Stanford, 1966-1968. Data, models, and evaluation of the arithmetic programs,
Academic Press, N.Y.

SUPPES, P. & ZINNES, J.L. (1963):
Basic measurement theory,
in: LUCE, R.D., BUSH, R.R. & GALANTER, E. (Hrsg.):
Handbook of mathematical psychology, I,
Wiley, N.Y.

TACK, W.H. (1970):
Messung als Repräsentation empirischer Gegebenheiten,
Zeitschrift für exp. und angew. Psychol., 17, 184-212

THISSEN, D. (1976):
Information in wrong responses to the Raven Progressive Matrices,
J. of Educ. Meas., 13, 201-214

THORNDIKE, R.L. (1982):
Test theory and test construction. Applied Psychmetrics,
Houghton & Mufflin, Dallas, Texas

THURSTONE, L.L. (1928):
The absolute zero in intelligence measurement,
Psychol. Rev., 35, 175-197

THURSTONE, L.L. (1947):
Multiple factor analysis,
The University of Chicago Press, Chicago, Ill.

THURSTONE, L.L. (1954):
An analytical method for simple structure,
Psychometrika, 19, 173-182

WAINER, H.U. & WRIGHT, B. (1980):
Robust estimation of ability in the Rasch model,
Psychometrica, 41, 3, 373-391

WARD, J. (1970):
On the concept of criterion-referenced measurement,
Brit. J. of Educ. Psychol., 40, 314-323

WARD, J. & FITZPATRICK, F. (1973):
Characteristics of matrices items,
Perceptual and Motor Skills, 36, 987-993

WEISS, D.J. (1982):
Improving measurement quality and efficiency with adaptive testing,
Applied Psychol. Meas., 6, 4, 473-492

WEISS, D.J. & DAVISON, M.L. (1981):
Test theory and methods,
Annual Rev. of Psychol., 32, 629-658

WHITE, A.P. & ZAMARELLI, J.E. (1981):
Convergent principles: Information in the answer sets of some multiple choice intelligence tests,
Applied Psychol. Meas., 5, 21-27

WILLMOT, A.S. & FOWLES, D.E. (1974):
The objective interpretation of test performance: The Rasch model applied,
Slough: National Foundation for Educational Research

WOLLENBERG, A.L. VAN DER (1979):
The Rasch model and time limit tests,
Doctoral Diss., Nijmegen

WOOD, R. (1973):
Response contingent testing,
Rev. of Educ. Res., 43, 529-544

WRIGHT, B.D. (1968):
Sample free test calibration and person measurement,
Proceedings of the 1967 Invitational Conference on Testing Problems,
Educational Testing Service, Princeton, N.J.

WRIGHT, B.D. (1977):
Solving measurement problems with the Rasch model,
J. of Educ. Meas., 14, 97-115

WRIGHT, B.D. & BELL, S.R. (1984):
Item banks: What, why, how?,
J. of Educ. Meas., 21, 4, 331-346

WRIGHT, B.D. & DOUGLAS, G.A. (1975):
Better procedures for sample-free item analysis,
Res. Memo. Nr.20, Statistical Laboratory, Dept. of Educ., University of Chicago

WRIGHT, B.D. & DOUGLAS, G.A. (1977):
Conditional vs. unconditional procedures for sample-free item analysis,
J. of Educ. and Psychol. Meas., 37, 573-586

WRIGHT, B.D., MEAD, R.J. & DRABA, R.E. (1976):
Detecting and correcting test item bias with a logistic response model,
Res. Memo. Nr.22, Statistical Laboratory, Dept. of Educ., University of Chicago

WRIGHT, B.D. & PANCHAPAKESAN, N. (1969):
A procedure for sample-free item analysis,
J. of Educ. and Psychol. Meas., 29, 23-48

WRIGHT, B.D. & STONE, M.H. (1979):
Best test design,
Mesa-Press, Chicago

ANHANG 1-5

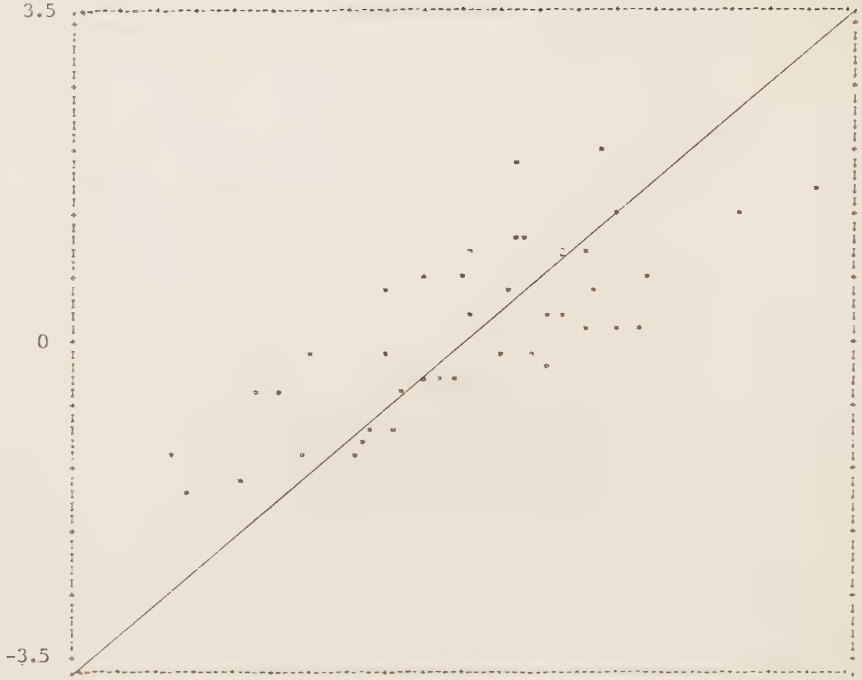
Test	Realisierungsmodi			geltende Regeln									Richtung der Regelgeltung			Distraktorenschlüssel			Richtigantworten							
	SC	JC	IR	I	A	Su	D	FA	Se	Vg	Vo		ws	w	s	1	2	3	A	B	C	D	E	I	G	H
1	10	6	8	6	10	8	8	2	3	5	6		11	21	16	6	7	11	4	0	4	6	3	4	1	2
2	8	8	8	4	7	7	7	7	6	5	5		13	16	19	8	10	6	4	0	1	5	1	9	2	2
3	8	8	8	3	5	5	7	6	6	7	8		21	13	14	10	10	4	0	6	4	5	0	6	0	5
4	8	8	8	7	8	5	6	5	4	6	6		18	16	14	6	12	6	2	3	5	2	3	4	1	4
5	8	8	8	6	8	7	5	7	5	5	4		17	13	18	9	5	10	7	2	4	3	0	2	3	3
6	8	8	8	6	9	7	4	7	5	5	5		20	13	15	9	8	7	4	3	3	5	0	4	3	2
7	8	8	8	6	5	8	6	5	6	6	6		15	17	16	9	6	9	5	1	2	4	1	4	4	3
8	8	8	8	7	8	6	7	6	6	3	5		17	13	18	5	10	9	0	4	2	1	2	6	4	5
9	8	8	8	6	12	6	4	6	8	1	5		19	14	15	10	12	2	1	2	5	4	1	5	4	2
10	8	8	8	5	11	4	6	4	8	3	7		15	19	14	12	7	5	4	1	3	1	3	6	2	4
11	8	8	8	5	5	5	7	4	5	7	10		17	15	16	9	9	6	1	2	3	5	2	6	2	3
12	8	8	8	4	9	8	5	6	5	5	6		20	14	14	11	5	8	0	3	4	5	2	3	3	4
13	8	8	8	8	4	8	7	7	5	4	5		11	22	15	5	9	10	1	4	2	4	6	3	1	3
14	8	8	8	9	5	9	5	4	6	7	3		15	10	23	10	5	9	0	2	4	6	3	4	2	3
15	8	8	8	6	11	7	3	5	9	5	2		17	16	15	8	11	5	2	1	4	4	3	3	2	5
16	8	8	8	6	10	10	5	4	3	5	5		15	18	15	9	8	7	1	1	5	6	1	5	2	3
17	8	8	8	7	9	8	5	6	4	3	6		18	14	16	11	5	8	3	4	4	5	2	4	2	0
18	8	8	8	5	7	3	7	8	7	6	5		17	20	11	7	10	7	5	8	3	4	1	2	0	1
19	8	8	8	5	4	7	7	11	4	5	5		13	20	15	9	5	10	1	2	5	6	3	3	3	1
20	8	8	8	6	7	6	7	9	4	4	5		15	15	18	7	8	9	0	2	3	8	2	2	1	6
21	9	8	7	14	5	6	9	4	6	2	2		17	15	16	12	5	7	1	5	1	2	0	5	4	6
22	8	8	8	4	6	8	9	7	8	3	3		18	15	15	10	6	8	3	6	1	3	2	3	4	2
23	8	8	8	3	6	10	7	6	8	5	3		21	12	15	8	8	8	1	4	6	3	2	3	2	3
24	8	8	8	4	12	3	5	10	7	4	3		17	19	12	7	9	8	1	6	7	1	3	1	2	3
25	8	8	8	5	4	5	6	6	9	6	7		13	18	17	10	10	4	1	1	6	2	1	4	2	2
26	8	8	8	3	5	5	6	10	9	5	5		13	17	18	11	5	8	5	4	2	2	2	2	3	4
27	8	8	8	4	7	3	4	11	14	3	2		14	17	17	6	10	8	3	3	5	3	3	2	3	2
28	8	8	8	6	4	4	7	9	5	6	7		15	16	17	9	9	6	2	5	4	5	4	2	1	1
29	8	8	8	3	3	5	5	7	10	8	7		20	13	15	9	10	5	2	5	2	5	3	4	1	2
30	8	8	8	4	9	3	3	7	13	6	3		21	14	13	7	9	8	1	5	1	5	1	2	2	7
31	8	8	8	4	8	3	5	5	9	9	5		14	16	18	9	6	9	3	2	2	4	3	3	2	5
32	8	8	8	2	5	3	3	7	13	9	6		15	17	16	9	7	8	3	3	1	3	4	5	1	4
33	8	8	8	6	7	3	1	9	11	7	4		19	15	14	9	10	5	5	6	2	3	2	3	1	2
34	8	8	8	3	6	3	3	9	8	9	7		18	16	14	10	8	6	1	3	2	4	5	0	5	4
35	10	6	8	2	4	3	4	10	10	8	7		17	16	15	11	10	3	5	2	3	0	3	1	5	5
	285	278	279	185	245	203	196	237	249	187	180		576	555	549	307	284	249	82	116	115	134	77	125	80	111

Anhang 1:
Verteilung der RM, RE, RR, Distraktorenschlüssel und Richtigantwortpositionen über die 35 Tests

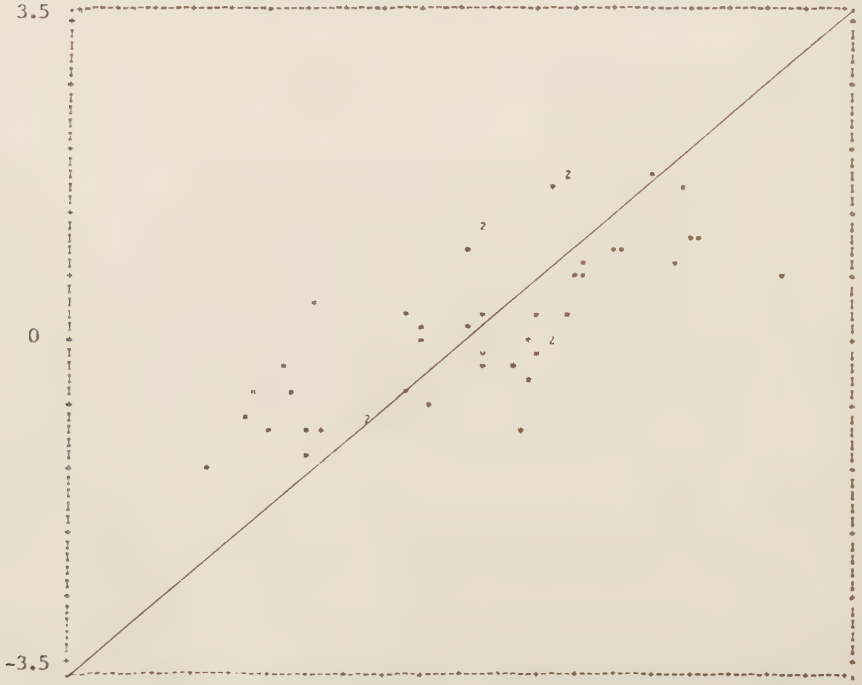
	TESTVERSION																																		
ROHWERT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	0	1	1	2	0	0	0	0	0	6	4	0	2	0	1	2	1	4	1	0	1	4	6	4	1	3	3	2	3	0	3	0	0	1	1
2	1	3	5	4	4	9	3	1	0	10	9	8	5	2	6	4	7	3	6	4	4	5	6	3	9	6	3	10	2	2	1	5	4	3	3
3	1	3	7	7	6	3	6	4	5	13	19	8	6	3	8	13	12	3	8	5	4	4	17	9	6	6	4	9	0	0	2	6	9	7	8
4	1	6	9	12	11	9	12	4	12	13	18	18	17	5	15	8	16	5	15	9	8	10	24	10	16	9	4	10	5	5	8	5	9	10	9
5	8	9	11	19	12	6	16	6	10	12	25	29	11	14	20	16	13	13	17	10	16	22	36	24	20	22	8	19	3	8	11	10	11	9	17
6	4	15	18	25	38	14	12	9	10	18	27	32	28	14	18	22	27	12	21	17	23	16	41	18	35	25	14	13	10	8	15	15	12	24	21
7	3	11	16	21	32	7	25	5	14	18	25	25	23	35	20	35	23	13	26	17	14	14	27	30	34	25	19	18	7	13	27	23	13	33	21
8	11	17	15	25	25	18	19	15	14	13	23	27	29	32	23	21	21	30	24	19	19	17	18	21	27	30	19	17	23	15	26	16	22	24	28
9	7	10	27	27	19	15	22	18	18	15	23	17	29	28	21	14	24	26	20	29	22	27	21	18	20	22	21	15	19	23	26	31	24	21	21
10	5	13	18	23	21	17	30	15	16	17	15	17	22	24	19	22	21	30	16	22	20	21	8	17	20	23	18	12	32	27	20	30	33	18	23
11	3	18	17	18	16	17	11	25	23	16	12	11	25	22	16	11	12	22	17	22	20	17	7	12	11	16	15	15	25	24	16	21	18	11	17
12	10	20	18	11	7	22	15	27	23	17	8	4	13	9	13	13	12	17	13	22	17	18	8	11	4	3	17	14	19	26	8	16	15	5	20
13	23	18	17	8	0	15	15	17	20	12	6	2	6	9	13	6	9	13	8	7	11	13	4	8	2	4	16	8	17	21	9	17	14	5	16
14	22	15	13	6	3	18	5	17	12	17	4	2	4	9	10	7	4	11	5	6	15	6	1	7	4	1	10	9	13	14	4	16	15	0	10
15	24	15	11	8	0	12	4	15	10	7	1	2	6	12	8	1	3	5	0	7	8	1	1	2	0	3	12	11	11	17	1	11	10	2	4
16	17	17	9	2	2	8	1	12	5	7	1	1	4	5	3	1	5	1	0	3	6	3	2	1	1	1	15	8	10	10	0	6	9	1	4
17	22	16	6	0	0	5	1	6	3	6	1	0	1	1	4	3	2	6	0	1	1	3	0	1	0	0	6	7	6	4	0	6	8	0	2
18	15	4	5	1	0	1	2	2	1	4	0	0	2	0	5	1	0	2	0	0	8	0	0	0	0	0	5	1	4	0	0	4	4	0	1
19	14	1	1	0	0	2	1	2	0	1	0	0	0	1	2	0	0	1	0	0	5	0	0	1	0	0	1	1	0	1	0	0	3	0	0
20	9	7	2	0	0	0	1	3	0	1	0	0	1	1	2	0	0	1	0	0	1	0	0	0	0	0	2	0	0	0	0	0	1	0	0
21	2	2	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	1	0	0
22	3	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Anhang 2:
Verteilungen der Rohwerte in den 35 Testversionen
(ohne Rohwerte 0 und 24)

Anhang 3
Abbildungen 3/1-12:
Gegenüberstellung der Schwierigkeitsparameter-
schätzungen im RM (Abzissen)
und der "zusammengesetzten" Itemschwierigkeiten
(Ordinaten) bei 12 Teildatensätzen
(vgl. Tabelle 9/Anhang 4)



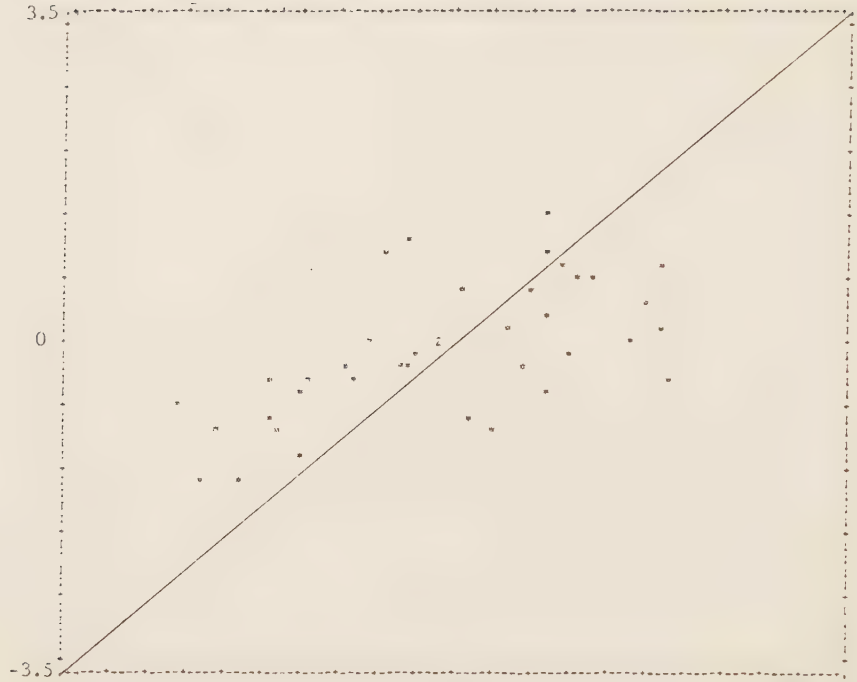
3/1



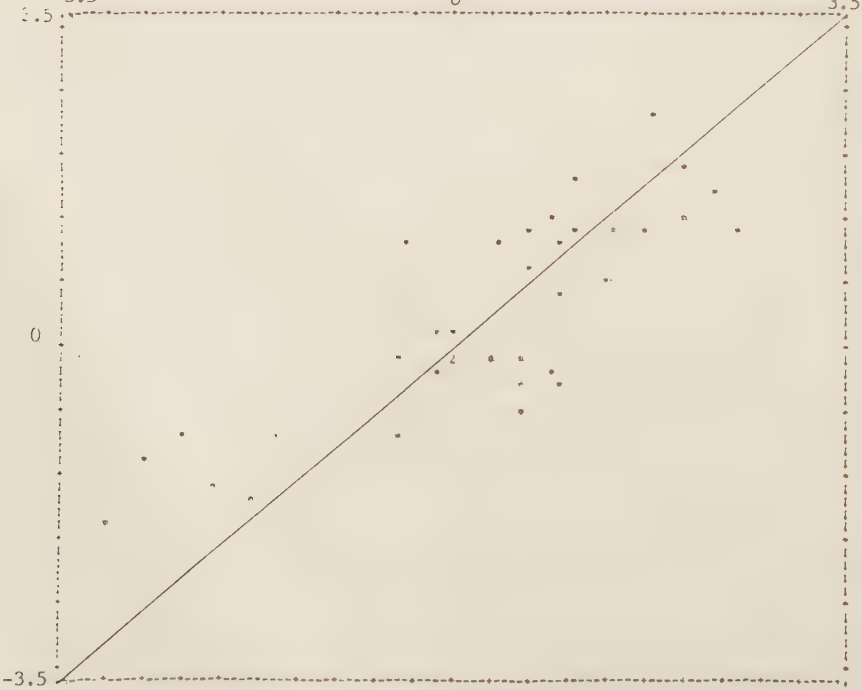
3/2

3/3

3/4



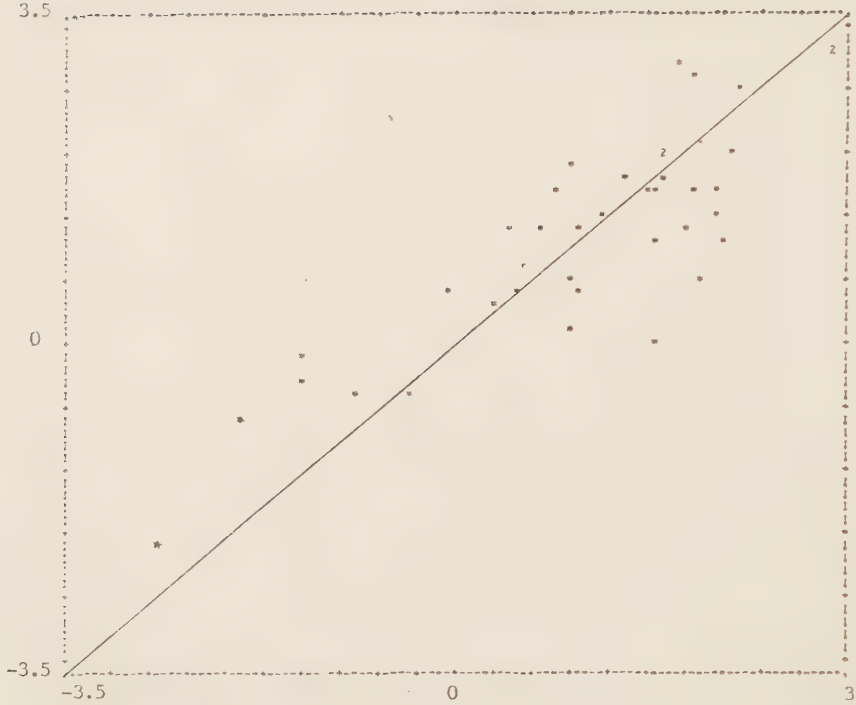
3/5



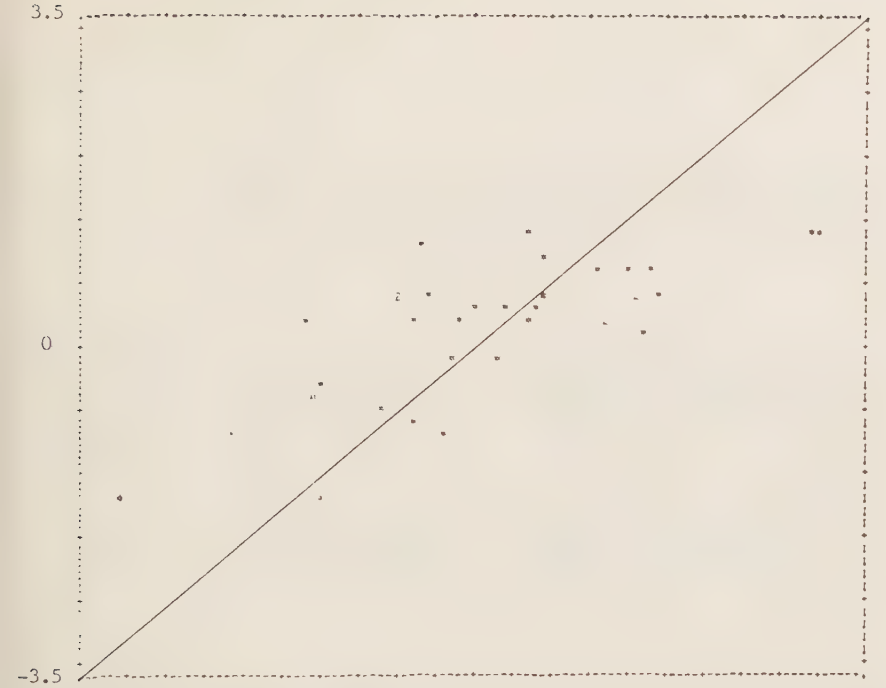
3/6

3/7

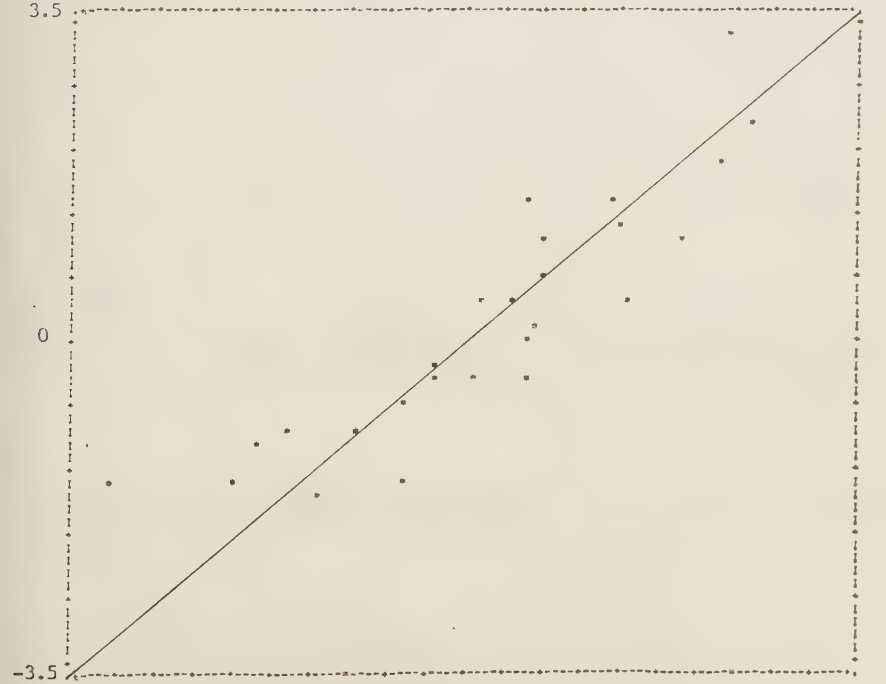
3/8


3.5
0
-3.5
-3.5
0
3.5
3/9

3.5
0
-3.5
3/10



3/11



3/12

Set:	I (k=43)	II (k=48)	III (k=38)	IV (k=41)	V (k=39)	VI (k=36)
SC	-0.27/0.06	-0.26/0.05	-0.97/0.07	-0.78/0.06	-0.35/0.06	-0.27/0.08
IC	-0.18/ --	-0.34/ --	-0.26/ --	-0.21/ --	-0.32/ --	-0.51/ --
IR	0.51/0.07	0.62/0.06	0.32/0.10	0.57/0.08	0.77/0.07	1.48/0.09
I	-0.88/0.08	-1.09/0.07	0.13/0.10	-1.65/0.09	-0.76/0.06	-0.43/0.11
A	0.02/0.06	0.17/0.06	-1.04/0.09	0.13/0.07	0.48/0.07	0.70/0.08
Su	-0.03/0.07	0.24/0.08	0.01/0.08	1.25/0.09	0.63/0.07	0.31/0.08
D	0.77/0.05	0.66/0.07	0.96/0.11	0.65/0.08	0.57/0.07	0.82/0.09
Ea	1.25/0.06	0.42/0.06	0.66/0.09	1.23/0.07	0.70/0.07	0.91/0.11
Se	-0.39/0.05	0.87/0.07	-0.19/0.07	-0.17/0.07	-0.12/0.06	-0.27/0.10
V_g	-0.18/ --	-0.34/ --	-0.26/ --	-0.21/ --	-0.32/ --	-0.51/ --
V_o	0.38/0.04	0.10/0.08	0.71/0.10	-0.10/0.06	-0.35/0.08	-0.33/0.11
w/s	-0.51/0.03	-0.53/0.03	-0.21/0.04	-0.44/0.04	-0.55/0.05	-1.01/0.06
w	-0.18/ --	-0.34/ --	-0.26/ --	-0.21/ --	-0.32/ --	-0.51/ --
s	-0.34/0.04	-0.19/0.03	-0.09/0.04	-0.12/0.04	-0.08/0.04	-0.37/0.05
	SD(n)=0.56	SD(n)=0.54	SD(n)=0.57	SD(n)=0.76	SD(n)=0.52	SD(n)=0.72

Set:	VII (k=39)	VIII (k=41)	IX (k=40)	X (k=39)	XI (k=34)	XII (k=27)
SC	-0.40/0.07	-0.18/0.06	-1.05/0.07	-0.82/0.07	-0.32/0.07	-0.25/0.10
IC	0.20/ --	-0.32/ --	0.09/ --	-0.31/ --	0.00/ --	-0.32/ --
IR	1.02/0.08	0.92/0.09	0.88/0.07	1.01/0.09	0.42/0.08	0.62/0.10
I	-0.79/0.08	-0.31/0.10	-0.92/0.09	-0.72/0.08	-1.00/0.10	-1.41/0.11
A	-0.66/0.09	-0.28/0.08	-0.57/0.09	-0.26/0.08	-0.06/0.07	-0.20/0.08
Su	0.07/0.08	-0.26/0.08	0.76/0.12	-0.13/0.08	-- / --	-- / --
D	0.80/0.08	-0.25/0.08	0.55/0.11	1.29/0.09	0.58/0.08	2.07/0.18
Ea	-0.02/0.09	0.85/0.08	1.13/0.10	0.75/0.06	0.28/0.06	-0.02/0.09
Se	-0.45/0.08	-0.13/0.07	-0.56/0.08	-0.34/0.07	-0.40/0.06	-0.71/0.06
V_g	0.20/ --	-0.32/ --	0.09/ --	-0.31/ --	0.00/ --	-0.32/ --
V_o	0.02/0.08	0.07/0.09	-0.08/0.08	0.63/0.07	0.21/0.07	0.02/0.08
w/s	-0.27/0.04	-0.04/0.04	-0.60/0.06	-0.58/0.05	-0.64/0.06	0.01/0.07
w	0.20/ --	-0.32/ --	0.09/ --	-0.31/ --	0.00/ --	-0.32/ --
s	0.13/0.05	0.57/0.04	0.13/0.05	0.10/0.04	-0.35/0.05	0.79/0.08
	SD(n)=0.51	SD(n)=0.44	SD(n)=0.68	SD(n)=0.66	SD(n)=0.45	SD(n)=0.83

Anhang 4
Tabelle 9:

Produktnormierte CML - Operationsparameterschätzungen (n) in den Datensätzen I-XII und deren Standardabweichungen (=Schätzfehler) SF(n) in der Darstellung ((n)/SF(n)). Zusätzlich aufgeführt sind die Standardabweichungen SD(n) der 14 Operationsschätzungen für jeden Datensatz I-XII und die jeweils zugrundeliegende Anzahl der Items (k).

Set	I+II 1(k=88)	III+IV 2(k=75)	V+VI 3(k=72)	VII+VIII 4(k=78)	IX+X 5(k=75)	XI+XII 6(k=58)	6 Matrizen AM(m)/SD(m)	12 Matrizen AM(n)/SD(n)
SC	-0.28	-0.69	-0.39	-0.27	-0.49	-0.20	-0.39 0.18	-0.42 0.31
IC	-0.17	-0.17	-0.40	-0.09	-0.03	-0.13	-0.17 0.13	-0.16 0.20
IR	0.73	0.25	1.16	1.12	1.16	0.62	0.84 0.37	0.88 0.42
I	-1.23	-0.89	-0.81	-1.00	-1.09	-1.31	-1.06 0.19	-0.98 0.50
A	0.04	-0.13	0.66	-0.32	-0.80	0.04	-0.09 0.48	-0.19 0.61
Su	-0.18	-0.35	0.45	-0.12	0.27	--	0.01 0.33	0.27 0.51
D	0.76	0.75	0.67	0.26	0.88	1.35	0.78 0.35	0.82 0.49
Ea	0.85	0.93	0.83	0.61	0.60	0.24	0.68 0.25	0.75 0.40
Se	-0.24	-0.09	-0.23	-0.35	-0.57	-0.53	-0.32 0.20	-0.35 0.22
V_g	-0.17	-0.17	-0.40	-0.09	-0.03	-0.13	-0.17 0.13	-0.16 0.20
V_o	0.37	0.27	-0.43	0.13	0.39	0.28	0.17 0.31	0.16 0.42
w/s	-0.35	-0.05	-0.64	-0.29	-0.59	-0.35	-0.38 0.21	-0.45 0.37
w	-0.17	-0.17	-0.40	-0.09	-0.03	-0.13	-0.17 0.13	-0.16 0.20
s	-0.01	0.01	-0.14	0.49	0.27	0.23	0.14 0.23	0.09 0.39
R(m,b) =	0.73	0.63	0.71	0.69	0.78	0.73		

	SC	IC	IR	I	A	Su	D	Ea	Se	V_g	V_o	w/s	w	s
k=446	-0.51	-0.23	0.85	-1.13	-0.05	0.23	0.76	0.81	-0.20	-0.23	0.23	-0.35	-0.23	0.08
R(m,b) =	0.65													

Anhang 5
Tabelle 10:
"MR-" Gewichte (m) der 14 Operationen, geschätzt aus sechs Matrizen mit je k Items und der Gesamtmatrix mit k=446 Items. Zusätzlich angegeben sind die multiplen Korrelationen R(m,b) sowie die Mittelwerte und Standardabweichungen der Schätzer für jede Operation bei 6 und 12 zugrundeliegenden Matrizen.

Lebenslauf

Michael Werner Habon

12.01.1956	geboren in Duisburg
1962-1966	Kath. Volksschule an der Musfeldstr. in Duisburg
1966-1975	Neusprachliches Steinbart-Gymnasium in Duisburg: Abitur
1975-1976	Grundwehrdienst in Nordheim und Bad Segeberg
1976-1977	Grundstudium der Allgemeinen Sprachwissenschaft an der Universität Düsseldorf
1977-1979	Grundstudium der Erziehungswissenschaft an der Universität Düsseldorf
1979-1981	Hauptstudium der Erziehungswissenschaft an der Philipps - Universität Marburg
28.10.1981	Diplom - Hauptprüfung für Erziehungswissenschaftler (Nebenfach: Psychologie)
seit 10.1981	Promotionsstudium im Fach Erziehungswissenschaft an der Universität Düsseldorf
1982-1985	Wissenschaftlicher Mitarbeiter an der Universität Düsseldorf im Forschungs- und Entwicklungsprojekt "Computerunterstütztes und adaptives Testen im Psychologischen Dienst der Bundeswehr"
seit 01.1986	Systemingenieur bei der Dornier GmbH, Friedrichshafen

FOTODRUCK J. MAINZ GmbH
Neupforte 13, Nähe Markt, D-5100 Aachen
Telefon: (0241) 27305 und (0241) 38606